



Meta-Evaluation Methodology and Benchmark for Automatic Story Generation

Cyril Chhun

► To cite this version:

Cyril Chhun. Meta-Evaluation Methodology and Benchmark for Automatic Story Generation. Machine Learning [cs.LG]. Institut Polytechnique de Paris, 2024. English. ⟨NNT : 2024IPPAT030⟩. ⟨tel-04975504⟩

HAL Id: tel-04975504

<https://theses.hal.science/tel-04975504v1>

Submitted on 4 Mar 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization



Meta-Evaluation Methodology and Benchmark for Automatic Story Generation

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (ED IP
Paris)
Spécialité de doctorat: Informatique

Thèse présentée et soutenue à Palaiseau, le 19 novembre 2024, par

CYRIL CHHUN

Composition du Jury :

Benoît Favre Professeur, Aix-Marseille Université (LIS)	Président/Examineur
Smaranda Muresan Associate Professor, Columbia University (DSI)	Rapporteuse
Sophie Rosset Directrice de recherche, Université Paris-Saclay (LISN)	Examinatrice
Marine Carpuat Associate Professor, University of Maryland (CLIP)	Examinatrice
Fabian Suchanek Professeur, Télécom Paris (LTCL)	Directeur de thèse
Chloé Clavel Directrice de recherche, INRIA Paris (ALMAAnaCH)	Co-directrice de thèse

This page intentionally left blank.

Acknowledgments

*“ L'important n'est pas ce qu'on a
fait de nous mais ce que nous
faisons nous-même de ce qu'on a
fait de nous. ”*

Saint Genet, comédien et martyr,
Jean-Paul Sartre

The three years of this PhD have been a long and arduous, yet enriching, endeavour. In many regards, this period of my life has been similar to the *Fool's Journey*, which I studied on a whim after an unexpected encounter with tarot card reading in the summer of 2023. Much like the Fool, who meets multiple influential figures on his road to enlightenment, I have met many wonderful people whom I feel obliged to mention here.

First of all, I address my warmest thanks to my two thesis supervisors, Fabian and Chloé. You have always been there for me when I needed you, especially during the difficult times; thank you for your unwavering support. I feel honored and blessed to have been allowed to pursue this PhD under your guidance.

Next, I would like to thank Claire Gardent and Smaranda Muresan, for having accepted to review this thesis, as well as Sophie Rosset, Benoît Favre, and Marine Carpuat, for having accepted to be part of my thesis committee. Thank you for having showed interest in my research work. I also want to thank Claire Gardent, again, and Oana Balalau, for having been part of my mid-term defence committee. Your positive and constructive feedback has meant a lot to me. Finally, I am very grateful to Pierre Colombo, my co-author, for having suggested to me the idea of my first article, which prompted me to reorient my research topic towards automatic evaluation. Thank you for your keen insight.

Now is perhaps a good time to mention all the people from Télécom Paris.

I address my sincere gratitude to the DIG team for the pleasant moments we shared together: the games of “The Crew” during lunch break, the DIG seminars and their iconic sushi orders (especially when I did get mine), the memorable DIG retreats, etc. I also want to thank the Social Computing group and more generally the people I met on the 5th floor during my

rather long pilgrimage over there. Special thanks go to Simon C., for our many interesting discussions and our refreshing meetups in Paris; to Aina, for being my COLING partner and probably one of the gentlest people I've ever met; to Nils, for organizing the DIG seminars and simply for being a nice guy in general; to Jean-Louis, for introducing me to the academic world all those years back; to Antoine, for making me discover the wonders of French versification, to Zacchary and Thomas, for our intense and adrenaline-packed spikeball games, and to Rajaa, for always sharing Moroccan delicacies with us. And, of course, thanks to everyone else: Chadi, Lihu, Pierre-Henri, Émile, Gaël, Matthieu, Lucien, Arturo, Simon D., Marc J., Louis, Gabriel, Roman, Tiphaine, Yanzhu, Lorraine, Alisa, Maria, Samuel, Alicia, François, Tom, Weichen, Minh Huong, Hoang, Mehwish, Lanfang, Yiwen, Marc H., Mariam, Chaitanya, Nicoline, Fajrian, and all the others I may have missed.

J'aimerais également remercier les personnes que j'ai rencontrées grâce au théâtre, en particulier au Studio Muller, et qui m'ont permis de vivre de très beaux moments pendant ces trois ans.

Merci à Lucas, mon grand frère retrouvé; à Jocelyn D. et Roman, pour ce pique-nique des enfers; à Emma C., pour cette saucisse volante et ces moments de sadisme; à mes autres camarades d'Analyse-Action : David, Maëva, Esther, Victor R., Audrey, Margaux, Marion, Justin, Romain, Marie L., Julie, Valentine, Énola, Simon, Thibault, Emma M., Thomas, Charlotte H., pour m'avoir accompagné dans cette recherche de la vie; à la compagnie B612 : Mélanie B., Nicky, Justine M., Juliette, Mélanie D., Quentin et Constance, pour ces écritures de plateau complètement farfelues; à mes répliques de concours : Setti, Alexandre et Hélène; à Katia et Vincent, pour ces deux années de prépa; à Daniel, pour cette ouverture à la vie que tu m'as apportée; à Jason, mon formidable professeur d'interprétation; à Jocelyn Muller, pour nos entretiens toujours très riches et pour cette merveilleuse école que tu as fondée. Merci également à Charlène, Manon R., Yoann, Merlin, Florian S., Benjamin, Patrice, Charlotte B., Ingrid, Marie G., Grégoire, Virgile, Aristide, Mélissa O., Justine A., Anaïs, Erwan, et bien d'autres.

Je tiens aussi à remercier spécifiquement Randal Douc, à qui je dois très certainement d'avoir osé me lancer à la fois dans ce doctorat et dans le théâtre, et Mme Élisabeth Desrousseaux, pour m'avoir aidé à naviguer dans le brouillard lorsque je me sentais perdu.

À présent, j'aimerais remercier mes proches.

Merci à Sylvain, mon ami de toujours; à Bruno, mon âme sœur; à Victor P., mon chef d'état-major, à Jeanne, ma rencontre; à ma coterie du Knight : Sébastien G., Sébastien C., Suzie et Sophia, pour nos fascinantes aventures; à Lucas, Mathieu, Nicky, Gabin, David, Esther, Élise, Mélanie B., Merlin, Alexandre, Paul et Yaëlle, pour nos belles discussions; à mes amis du lycée : Louis, Marc et Erwan; à mes co de prépa : Loïc et Arnaud; à mes camarades de Coët : Moïse, François, Alexis et Antonin; à Amélie, ma partenaire de musées; à Éléonore, pour nos agréables sorties; à Adèle, pour cette enivrante soirée; à Camille, pour cette douce après-midi; enfin, à ceux que j'ai eu la chance d'apprendre à connaître : Antoine V.-L., Jocelyn D., Louis-Vincent

et Léa, Kristina, Tristan, Jean-Baptiste, Stépan, Thibault, Ruben P., Frédérique et Florian L., Samuel, Tarek, Antoine D., Romain L., Côme, Théodore...

Pour finir, je souhaite remercier ma famille.

Merci à Neak Mê, Pou Pin et Tata Anaïg, Pou Nin et Tata Pach, Mac Ming et Jean-Pierre. Merci à Om Srun et Mak Om, Om Kimly et Om Ny. Merci à mes cousins et cousines : Tana, Johan, Manon, Charlise, Matéo, Estelle, Alexandra, Amandine et Bang Kosal. Merci à mon frère Victor. Merci à ma mère, pour sa tendresse, et à mon père, pour son soutien.

À tous, je vous suis infiniment reconnaissant de tout ce que vous m'avez apporté.

Cela dit, pour paraphraser cette belle réplique de *Racine carrée du verbe être*, de Wajdi Mouawad :

“ Je ne vous aime pas parce que vous m'avez rendu meilleur ; je vous aime parce que je vous aime. ”

Vanves, le 3 décembre 2024.

A handwritten signature in black ink, appearing to read 'Chhun', with a long, sweeping horizontal stroke extending to the right.

Cyril Chhun

Abstract

Storytelling is a central component of human culture: it shapes societies through the transmission of shared traditions and common knowledge. Multiple approaches have been proposed to explore computational storytelling, despite the inherent challenges posed by the tasks of generating stories and assessing their quality. In this thesis, we design a meta-evaluation methodology and benchmark for Automatic Story Generation (ASG).

First, we lay the groundwork for conducting our meta-evaluation: we describe our chosen setting, provide definitions for the ASG and Automatic Story Evaluation (ASE) tasks, and review the ASG and social sciences literature to propose an original, comprehensive set of six criteria for story evaluation: Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity. We also review and select multiple automatic measures and group them w.r.t. a newly-proposed, two-dimensional taxonomy. Finally, we introduce our meta-evaluation framework, which we adapted specifically to ASG through the design of a new meta-evaluation strategy: the overall correlation.

Then, in order to use our human evaluation criteria and to conduct a meta-evaluation with our proposed framework, we introduce HANNA, our corpus of Human ANnotated NArratives. We begin by reviewing corpora related to ASG and find that no existing dataset contains human annotations w.r.t. specific criteria of story quality. We select several ASG systems to build HANNA, which contains 1,056 stories produced by 11 different systems, aligned on 96 prompts, and annotated w.r.t. our six criteria. We analyze the overall and system-level correlations between our criteria, showing that they allow for a standardized human evaluation. We also observe that large pre-trained language models such as GPT-2 produce the best results for ASG. We use Large Language Models (LLMs) to augment HANNA with 480 new stories and conduct an annotation experiment with LLMs which allows us to collect 150k+ annotations of story ratings w.r.t. different Eval-Prompts. Mainly, we observe that LLMs obtain better grades than humans, as rated by our selected LLMs: Beluga-13B and Mistral-7B.

After that, we perform our meta-evaluation benchmark of automatic measures on HANNA, where we study their overall and system-level correlations with human ratings through different strategies, which we complete with more fine-grained analyses: an experiment on the correlations of top- k systems, a pairwise system comparison with bootstrap resampling, and a discussion on the statistical discernibility of our results. We mainly observe that overall correlations of automatic measures with human criteria remain weak, meaning that specific measures for ASE are needed, and that commonly-used measures such as BLEU are sub-optimal compared with other existing measures (*e.g.* BERTSCORE and BARTSCORE). We then

show our analysis of LLM performance at ASE: we compare LLMs to human and non-LLM automatic evaluation methods and we examine the influence of the Eval-Prompt on LLM performance. We find that LLMs are currently the best proxy for human evaluation of ASG, especially w.r.t. system-level correlations, that LLMs are remarkably self-consistent, and that, in our specific setting, providing detailed guidelines does not improve correlations between LLM and human ratings.

Our results prompt us to study whether the performance displayed by LLMs at ASE and ASG can be explained through different factors. We perform two main experiments: a three-part study on LLM-generated explanations, and an analysis of pretraining data on LLM performance. For the former, (1) we compute contextual embeddings for each LLM explanation in order to perform a clustering analysis and show that explanations are well-separated w.r.t. our criteria, (2) we perform a keyword analysis that confirms the previous findings, and (3) we conduct a user study to assess the appropriateness of LLM explanations. Notably, we find that LLMs struggle to explain their answers with substantiated claims. For the latter, we use a detection method based on minimum token probabilities to show that larger models tend to produce content that is more similar to that of existing books.

To conclude, we outline three main research perspectives: designing specific ASE measures, further investigating LLM performance at ASG and ASE, and assessing and mitigating the impact of LLMs on society.

Résumé

La narration d’histoires est un élément central de la culture humaine : elle façonne les sociétés par la transmission de traditions partagées et de connaissances communes. De nombreuses approches de narration numérique ont été proposées, malgré les défis inhérents aux tâches de génération d’histoires et d’évaluation de leur qualité. Dans cette thèse, nous concevons une méthodologie et un benchmark pour la méta-évaluation de la génération automatique d’histoires (ASG).

Tout d’abord, nous posons les bases de notre méta-évaluation : nous décrivons le contexte choisi, définissons les tâches d’ASG et d’évaluation automatique d’histoires (ASE), et passons en revue les littératures d’ASG et des sciences sociales afin de proposer un ensemble original de six critères pour l’évaluation d’histoires : Pertinence, Cohérence, Empathie, Surprise, Engagement et Complexité. Nous examinons et sélectionnons également de multiples mesures automatiques et les regroupons dans une taxonomie bidimensionnelle nouvellement proposée. Enfin, nous présentons notre cadre de méta-évaluation, que nous avons adapté spécifiquement à l’ASG en concevant une nouvelle stratégie de méta-évaluation : la corrélation globale.

Ensuite, afin d’utiliser nos critères d’évaluation humaine et de mener une méta-évaluation avec le cadre proposé, nous présentons HANNA, notre corpus d’histoires annotées manuellement (Human ANnotated NArratives). Nous commençons par passer en revue les corpus liés à l’ASG et constatons qu’aucun ensemble de données existant ne contient d’annotations humaines en fonction de critères spécifiques de qualité des histoires. Nous sélectionnons plusieurs systèmes d’ASG pour construire HANNA, qui contient 1 056 histoires produites par 11 systèmes différents, alignées sur 96 prompts, et annotées selon nos six critères. Nous analysons les corrélations globales et au niveau du système entre nos critères pour montrer que ces derniers permettent une évaluation humaine standardisée. Nous observons également que les modèles linguistiques pré-entraînés de plus grande taille tels que GPT-2 produisent les meilleurs résultats pour l’ASG. Nous utilisons des grands modèles de langage (LLM) pour ajouter 480 nouvelles histoires à HANNA et nous menons une expérience d’annotation avec des LLM qui nous permet de collecter 150k+ annotations de notes d’histoires en fonction de différents prompts d’évaluation. Nous observons principalement que les LLM obtiennent de meilleures notes que les humains, tels que notés par les LLM que nous avons sélectionnés : Beluga-13B et Mistral-7B.

Ensuite, nous effectuons notre méta-évaluation comparative des mesures automatiques sur HANNA, où nous étudions leurs corrélations globales et au niveau du système avec

les évaluations humaines par le biais de différentes stratégies, que nous complétons avec des analyses plus minutieuses : une expérience sur les corrélations des systèmes top- k , une comparaison des systèmes par paires avec rééchantillonnage bootstrap, et une discussion sur la discernabilité statistique de nos résultats. Nous observons principalement que les corrélations globales des mesures automatiques avec les critères humains demeurent faibles, ce qui souligne la nécessité de concevoir des mesures spécifiques pour l'ASE, et que les mesures couramment utilisées telles que BLEU sont sous-optimales par rapport à d'autres mesures existantes comme BERTSCORE et BARTSCORE. Nous montrons ensuite notre analyse de la performance des LLM pour l'ASE : nous comparons les LLM aux méthodes d'évaluation automatique humaines et non-LLM et nous examinons l'influence du prompt d'évaluation sur la performance des LLM. Nous constatons que les LLM sont actuellement la meilleure approximation de l'évaluation humaine de l'ASG, en particulier en termes de corrélations au niveau du système, que les LLM sont remarquablement cohérents avec eux-mêmes, et que, dans notre contexte spécifique, fournir des consignes détaillées n'améliore pas les corrélations entre les LLM et les évaluations humaines.

Nos résultats nous incitent à étudier si les performances affichées par les LLM en ASE et ASG peuvent être expliquées par différents facteurs. Nous réalisons deux expériences principales : une étude en trois parties sur les explications générées par les LLM et une analyse des données de pré-entraînement sur les performances des LLM. Dans le premier cas, (1) nous construisons des plongements contextuels pour chaque explication de LLM afin d'effectuer une analyse de partitionnement : nous montrons ainsi que les explications sont correctement séparées en fonction de nos critères ; (2) nous effectuons une analyse de mots-clés qui confirme les résultats précédents, et (3) nous menons une étude utilisateur pour évaluer l'adéquation des explications LLM. Notamment, nous constatons que les LLM éprouvent des difficultés à expliquer leurs réponses par des affirmations étayées. Dans le second cas, nous utilisons une méthode de détection fondée sur les probabilités minimales de jeton pour montrer que les modèles plus importants ont tendance à produire un contenu plus similaire à celui des livres existants.

Pour conclure, nous présentons trois perspectives de recherche principales : la conception de mesures spécifiques à l'ASE, l'étude plus approfondie des performances des LLM en ASG et ASE, et l'évaluation et l'atténuation des risques des LLM sur la société.

Contents

Acknowledgments	i
Abstract	v
Résumé	vii
Contents	ix
List of Figures	xi
List of Tables	xv
1 Introduction and Overview	1
1.1 Introduction	2
1.2 Research Questions	4
1.3 Organization of This Thesis	5
1.4 Publications	7
2 Background	9
2.1 Pre-neural Automatic Story Generation	10
2.2 Language Models	14
2.3 Evaluation and Meta-Evaluation	32
2.4 Conclusion	34
3 Methodology for Story Evaluation and ASG Meta-Evaluation	37
3.1 Introduction and Definitions	38
3.2 Designing New Criteria for Human Story Evaluation	39
3.3 Selecting Automatic Measures for Story Evaluation	49
3.4 Proposed Framework for Meta-Evaluation	58
3.5 Conclusion	62
4 HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation	63
4.1 Introduction	64
4.2 Design Decisions for Building HANNA	65
4.3 The HANNA Corpus	74

4.4	Augmenting HANNA with Large Language Models	82
4.5	Conclusion	92
5	Meta-Evaluation Benchmark of Automatic Story Generation	95
5.1	Meta-Evaluation on HANNA V1	96
5.2	Large Language Models for Automatic Story Evaluation	111
5.3	Conclusion	118
6	Exploring LLM Explainability for ASE and ASG	121
6.1	Introduction	122
6.2	Related Work	123
6.3	Analysis of the LLM Explanations of our ASE Experiments	124
6.4	Influence of Pretraining Data on ASG Performance	129
6.5	Conclusion	131
7	Conclusion	133
7.1	Contributions	134
7.2	Perspectives	135
8	Bibliography	141
A	Appendix	I
A.1	The 96 HANNA Prompts	I
A.2	Supplementary Figures	VI

List of Figures

1.1	An introduction to this thesis generated by the Mistral language model with the following prompt: “Write the introduction of a PhD thesis whose topic is ‘Meta-Evaluation Methodology and Benchmark for Automatic Story Generation.’”	3
2.1	Structural story generation. Source: Alhussain and Azmi [2021] .	10
2.2	Plot graph for the sentence: “ <i>the princess loves the prince</i> ” [McIntyre and Lapata, 2010].	14
2.3	A fully-connected feed-forward neural network [Neutelings, 2021].	19
2.4	A layer of a convolutional neural network [Stutz, 2020].	19
2.5	A fully-recurrent neural network [User121799, 2019].	20
2.6	An LSTM cell [V, 2018].	21
2.7	CBOW and Skip-gram architectures [Mikolov et al., 2013a].	24
2.8	The transformer architecture [Vaswani et al., 2017].	27
2.9	Scaled dot-product attention [Vaswani et al., 2017].	27
2.10	Multi-head attention [Vaswani et al., 2017].	27
4.1	Overall absolute Kendall correlations between human criteria. Coefficient values are multiplied by 100 for readability; we will symbolize this with “(×100)” in the next figures.	81
4.2	System-level absolute Kendall correlations (×100) between human criteria.	81
4.3	Example Eval-Prompt and answer from our experiments. “Prompt” inside the Eval-Prompt refers to the story-prompt.	83
4.4	Schema of the performed ASE experiments. RE, CH, etc. are the considered human criteria. “EP” means “Eval-Prompt”, defined in sous-section 4.4.2. For the user study (sous-section 6.3.3), we randomly sampled 100 explanations from our experiments.	87
4.5	Example of an Eval-Prompt 1 for the Surprise criterion.	88
4.6	Example of an Eval-Prompt 2 for the Surprise criterion.	88
4.7	Example of an Eval-Prompt 3 for the Surprise criterion.	89
4.8	Example of an Eval-Prompt 4 for the Surprise criterion.	90
5.1	Overall absolute Kendall correlations (×100) between automatic measures and criteria. Full version shown on figure A.1.	97

5.2	System-level absolute Kendall correlations ($\times 100$) between automatic measures and criteria. Full version shown on figure A.2.	98
5.3	Overall absolute Kendall correlations ($\times 100$) between automatic measures. Full version shown on figure A.3.	100
5.4	System-level absolute Kendall correlations ($\times 100$) between automatic measures. Full version shown on figure A.4.	101
5.5	System-level absolute Kendall correlations between automatic measures and our proposed human criteria on top- k systems.	102
5.6	Weighted macro F1-scores ($\times 100$) of paired bootstrap resampling. Full version shown on figure A.5.	103
5.7	BH-adjusted p -values ($\times 100$) of the Relevance Williams tests for overall Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	104
5.8	BH-adjusted p -values ($\times 100$) of the Coherence Williams tests for overall Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	104
5.9	BH-adjusted p -values ($\times 100$) of the Empathy Williams tests for overall Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	105
5.10	BH-adjusted p -values ($\times 100$) of the Surprise Williams tests for overall Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	105
5.11	BH-adjusted p -values ($\times 100$) of the Engagememnt Williams tests for overall Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	106
5.12	BH-adjusted p -values ($\times 100$) of the Complexity Williams tests for overall Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	106
5.13	BH-adjusted p -values ($\times 100$) of the Relevance Williams tests for system-level Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	107
5.14	BH-adjusted p -values ($\times 100$) of the Coherence Williams tests for system-level Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	107
5.15	BH-adjusted p -values ($\times 100$) of the Empathy Williams tests for system-level Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	108
5.16	BH-adjusted p -values ($\times 100$) of the Surprise Williams tests for system-level Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	108
5.17	BH-adjusted p -values ($\times 100$) of the Engagememnt Williams tests for system-level Kendall correlations. Lower is better. "0" means $p < 0.01$. Blank cells mean there is a decrease in correlation.	109

5.18	BH-adjusted p -values ($\times 100$) of the Complexity Williams tests for system-level Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.	109
5.19	Overall absolute Kendall correlations between evaluation measures and human ratings. Higher is better. The black vertical line separates LLMs (left) and non-LLMs (right). Coefficient values are multiplied by 100 for readability; we will symbolize this with “($\times 100$)” in the next figures.	113
5.20	System-level absolute Kendall correlations ($\times 100$) between evaluation measures and human ratings. Higher is better. The white vertical line separates LLMs (left) and non-LLMs (right).	113
5.21	BH-adjusted p -values ($\times 100$) of the Williams tests for overall and system-level Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.	115
5.22	Overall absolute Kendall correlations ($\times 100$) between LLMs and human ratings for different Eval-Prompts. Higher is better. B-13B = Beluga-13B, L-13B = Llama-13B, M-7B = Mistral-7B and Chat = ChatGPT.	117
5.23	System-level absolute Kendall correlations ($\times 100$) between LLMs and human ratings for different Eval-Prompts. Higher is better. B-13B = Beluga-13B, L-13B = Llama-13B, M-7B = Mistral-7B and Chat = ChatGPT.	118
6.1	UMAP projection of Beluga-13B explanations.	124
6.2	UMAP projection of Mistral-7B explanations.	125
A.1	Overall absolute Kendall correlations ($\times 100$) between automatic measures and criteria.	VII
A.2	System-level absolute Kendall correlations ($\times 100$) between automatic measures and criteria.	VIII
A.3	Overall absolute Kendall correlations ($\times 100$) between automatic measures. . .	IX
A.4	System-level absolute Kendall correlations ($\times 100$) between automatic measures. . .	X
A.5	Weighted macro F1-scores ($\times 100$) of paired bootstrap resampling.	XI

List of Tables

2.1	The only surviving output from Joseph Grimes’ system [Ryan, 2017].	11
3.1	Overview of human story evaluation practices in ASG (Part 1 of 2).	44
3.2	Overview of human story evaluation practices in ASG (Part 2 of 2).	45
3.3	Taxonomy of the automatic measures considered in our study with symbols for easier identification. Note: BARTSCORE was designed to be either reference-based or reference-free depending on the setting.	50
3.4	Gradual evidence language proposal [Muff et al., 2022].	62
4.1	Example stories from ROCSTORIES [Mostafazadeh et al., 2016].	65
4.2	Example stories from the Story Cloze Test [Mostafazadeh et al., 2016].	66
4.3	Example story from SIND [Huang et al., 2016].	66
4.4	Example stories from WRITINGPROMPTS [Fan et al., 2018].	67
4.5	Example descriptions from ROLEPLAYERGUILD [Louis and Sutton, 2018].	68
4.6	Overview of existing story generation corpora. No corpus provides annotations on different criteria of story quality.	69
4.7	Example aligned stories from FUSION, HINT, and TD-VAE.	75
4.8	Example task from our Amazon Mechanical Turk experiment.	78
4.9	Example prompt, human and generated stories from HANNA with human annotations and measure scores.	79
4.10	Intra-class coefficient for average random raters per criterion. “95% LB” and “95% UB” stand for lower- and upper-bounds of the 95% confidence interval respectively. Higher is better.	80
4.11	Average system ratings per criterion with 95% confidence interval. Best value in bold blue , values in the confidence interval of the best value in bold without asterisk. Higher is better.	82
4.12	Example aligned stories from Llama-7B, Beluga-13B, and Platypus2-70B.	84
4.13	Average Beluga-13B ratings for Eval-Prompt 1 with 95% confidence interval. Higher is better.	91
4.14	Average Mistral-7B ratings for Eval-Prompt 1 with 95% confidence interval. Higher is better.	92

5.1	Top 3 measures per criterion for overall Kendall (τ) correlation values. Indices denote different variants.	99
5.2	Top 3 measures per criterion for system-level Kendall (τ) correlation values. Indices denote different variants.	99
5.3	Top 5 measures computed by one-level ranking per aggregation level, higher Borda count (BC) is better	110
5.4	Intra-class coefficients type 2k for Eval-Prompt 1 ratings with 95% confidence interval. Higher is better.	111
5.5	Intra-class coefficients type 2k for Beluga-13B ratings with 95% confidence interval. Higher is better.	116
5.6	Average Likert ratings per LLM per Eval-Prompt. The asterisk signals the fact that ChatGPT was only asked to rate the HANNA V1 stories. Higher is better. . .	116
6.1	Adjusted Rand index and adjusted mutual information for Beluga-13B and Mistral-7B clusterings. Higher is better.	125
6.2	Selected keywords from Beluga-13B explanations w.r.t. a specific criterion. . . .	126
6.3	Example task from our Amazon Mechanical Turk user study.	128
6.4	Error rates of Beluga-13B Eval-Prompt 3 on a sample of 100 explanations. Lower is better. The asterisk signals that all 1,056 Eval-Prompt 3 annotations were considered.	129
6.5	Predicted contamination rates of the WRITINGPROMPTS sample.	130
6.6	AUC detection score on the BOOKSMIA dataset.	130

Chapter 1

Introduction and Overview

*“ I never quite realized... how
beautiful this world is. ”*

A2 (NieR:Automata)

Contents

1.1 Introduction	2
1.2 Research Questions	4
1.2.1 RQ1: How Can We Evaluate the Quality of a Story?	4
1.2.2 RQ2: How Good Are Existing Models at Generating Stories?	4
1.2.3 RQ3: To Which Extent Can We Use Automatic Measures for Story Eval- uation?	4
1.2.4 RQ4: How Explainable Are the Evaluation Ratings of Large Language Models?	4
1.3 Organization of This Thesis	5
1.4 Publications	7

1.1 Introduction

Creativity is a long-standing goal of research in Artificial Intelligence (AI). The *Proposal* [McCarthy et al., 2006] for the 1956 Dartmouth Summer Research Project on Artificial Intelligence—widely considered to be the founding event of AI as a field—lists creativity as one of its seven aspects of the “AI problem”, surmising that the difference between creative thinking and unimaginative competent thinking lies in the injection of randomness. Indeed, creativity can be defined as “the ability to generate novel, and valuable, ideas” [Boden, 2009], where *valuable* can be understood as interesting, beautiful, useful, meaningful, elegant, etc. It is a complex process that involves exploring different possibilities and taking risks; it requires intuition, imagination, and emotional intelligence. It is therefore often argued that creativity is specific to human nature and that no computer could ever exhibit “true” creativity. Nevertheless, this has not deterred AI researchers from studying how computers could perform creative tasks. In particular, the automatic generation of stories has been a staple of the arduous quest for artificial creativity.

Stories have played an integral role in shaping human history and societies throughout time [Sinclair, 2005]. Oral traditions, myths, and legends have been passed down from generation to generation, providing a link between past and present. Stories are often used as a way to teach moral lessons, impart wisdom, and pass on practical skills. For example, many indigenous cultures use stories to teach children about the natural world, how to hunt and fish, and how to live sustainably. Similarly, folktales have been used throughout history to teach children important life skills such as respect for authority, honesty, and perseverance. In addition to preserving cultural heritage [Paolini and Di Blas, 2014] and transmitting knowledge [Kreiwirth, 2000], stories provide entertainment and inspiration for people of all ages. They can be a source of joy, laughter, and emotional catharsis, helping individuals cope with the stresses and challenges of daily life.

In recent years, research in cognitive sciences has shown that skillful storytelling enables a narrator to build an authentic connection with their audience and allows listeners to grasp complex concepts more intuitively [Suzuki et al., 2018]. Storytelling-based approaches are now used in multiple fields, such as:

1. Communication: storytelling serves as a common ground for establishing a value-laden reality among diverse parties and help build a social relationship more quickly [Barker and Gower, 2010];
2. Education: stories can become a tool for asserting a life-enhancing value system, where knowledge can be shared in the public interest [Woodhouse, 2011];
3. Gaming: storytelling allows for an experience that is modified in real time and tailored to the preferences of the player in order to provide a more personalized end result [Bostan and Marsh, 2010];
4. Marketing: well-crafted storytelling may create positive associations with a consumer brand and increase consumers’ willingness to pay for it [Lundqvist et al., 2013].

Since the 1960s, multiple approaches have been used to explore computational storytelling. Research in the field involves different aspects, *e.g.* story understanding, story representation and story generation [Alabdulkarim et al., 2021]. Furthermore, the recent development of transformer-based language models [Devlin et al., 2019; Vaswani et al., 2017] has led to the emergence of Large Language Models (LLM) [Brown et al., 2020], which have demonstrated impressive performance at many Natural Language Processing tasks. Figure 1.1 shows an introduction to this thesis that has been automatically generated by an LLM called Mistral (more precisely, Mistral-7B-Instruct-v0.1-GGUF) and eloquently illustrates current language model capabilities. Most notably, Mistral produces text with impeccable fluency and, since it was given only the basic topic of the thesis, makes reasonable claims about its contents.

Automatic story generation (ASG) has seen rapid advancements in recent years, with large language models (LLMs) like BERT, RoBERTa, and GPT-3 demonstrating impressive performance. However, evaluating the quality of generated stories remains a challenging task due to subjectivity and lack of standardization. This calls for the development of robust meta-evaluation methodologies and benchmarks that can accurately assess ASG models' performance across various metrics.

In this PhD thesis, we propose a comprehensive meta-evaluation framework for automatic story generation. Our approach combines multiple evaluation methods, including human assessment, expert judgment, and automated metrics, to provide a holistic understanding of model performance. We also introduce a novel benchmark dataset specifically designed for ASG, which includes diverse genres and narrative structures, enabling us to evaluate models' ability to generate coherent and engaging stories across various contexts.

Our contributions include:

1. A comprehensive meta-evaluation framework that integrates multiple evaluation methods, providing a more accurate assessment of ASG model performance.
2. A novel benchmark dataset specifically designed for automatic story generation, enabling us to evaluate models' ability to generate coherent and engaging stories across various contexts.
3. An in-depth analysis of the strengths and weaknesses of different LLMs, shedding light on their suitability for specific ASG tasks and genres.
4. A discussion on the importance of transparency and reproducibility in ASG research, emphasizing the need for open-source models and standardized evaluation methodologies.

Through our work, we aim to advance the field of automatic story generation by providing a robust and comprehensive framework for model evaluation, enabling researchers to compare and improve their models more effectively.

Figure 1.1: An introduction to this thesis generated by the Mistral language model with the following prompt: "Write the introduction of a PhD thesis whose topic is 'Meta-Evaluation Methodology and Benchmark for Automatic Story Generation'."

This setting, *i.e.*, "given a prompt, generate a story," will be the one used throughout our work.

1.2 Research Questions

In this section, we outline the four research questions that we address in this thesis.

1.2.1 RQ1: How Can We Evaluate the Quality of a Story?

While stories are central in human culture, determining whether a story is good or bad is fairly difficult. Compared to other Natural Language Processing (NLP) tasks such as machine translation and text summarization, for which specific practices for evaluation have been designed, story evaluation is still an open question, especially due to its subjective nature. In [Chapter 3](#), we propose a methodology for story evaluation, including a comprehensive set of orthogonal criteria.

1.2.2 RQ2: How Good Are Existing Models at Generating Stories?

There have been many approaches to Automatic Story Generation (ASG), ever since the first systems were proposed in the 1960s [[Propp, 1968](#); [Ryan, 2017](#)]. However, many of those models produced stories that were still far from human performance. Recently, the development of neural networks has enabled language models to demonstrate impressive performance at many NLP tasks, among which ASG. This naturally begs the question of whether automatically generated stories can compete with human-written ones. In [Chapter 4](#), we design HANNA, a corpus of stories annotated with both human and LLM ratings, which allows us to compare existing systems with one another and with human writers.

1.2.3 RQ3: To Which Extent Can We Use Automatic Measures for Story Evaluation?

Human evaluation has always been the gold standard for the evaluation of any task in NLP, but it is costly and time-consuming. To answer those issues, automatic evaluation measures were developed, such as BLEU [[Papineni et al., 2002](#)] for machine translation. However, since there have been no automatic measure specifically designed for Automatic Story Evaluation (ASE), researchers have been relying on measures that were introduced for other tasks. In [Chapter 5](#), we perform an extensive meta-evaluation of ASG to assess whether existing automatic measures are appropriate for ASE, going from simple textual measures to Large Language Models used as automatic annotators.

1.2.4 RQ4: How Explainable Are the Evaluation Ratings of Large Language Models?

We find in [Chapter 5](#) that LLMs outperform other evaluation measures for ASE, and that they are currently the best proxy for human evaluation of stories. However LLMs are well-known for behaving in ways that are difficult to understand, and, as their usage becomes ever more prevalent, it is becoming increasingly important to assess how reliable they are. In [Chapter 6](#), we perform two experiments that aim at providing a better understanding of LLM performance at ASE and ASG: a three-part study on LLM-generated explanations, and an analysis of the influence of pretraining data on their ASG performance.

1.3 Organization of This Thesis

In this section, we provide a summary of each chapter of this thesis.

Chapter 2: Background. First, we provide some background information related to Automatic Story Generation and Evaluation (ASG and ASE). In [Section 2.1](#), we cover older methods for ASG: namely, structural and planning-based models. In [Section 2.2](#), we present language models, focusing on neural-based language models, from recurrent neural networks to large language models. Finally, in [Section 2.3](#), we present different aspects related to the question of evaluation in NLP.

Chapter 3: Methodology for Story Evaluation and ASG Meta-Evaluation. In this chapter, we lay the groundwork for conducting a meta-evaluation of Automatic Story Generation. We begin by describing our chosen setting and providing definitions for the ASG and ASE tasks ([Section 3.1](#)). Then, we review the human evaluation methods used in the ASG literature ([Section 3.2.1](#)) and several approaches from the social sciences literature ([Section 3.2.2](#)). This allows us to propose an original, comprehensive set of six criteria ([Section 3.2.3](#)). We select and review a number of automatic evaluation measures designed for diverse NLP tasks ([Section 3.3.2](#)) and grouped them in different classes w.r.t. a newly-proposed taxonomy ([Section 3.3.1](#)). Finally, we describe our meta-evaluation framework, which we tailored for the ASG task ([Section 3.4](#)).

Chapter 4: HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation. In order to use our human evaluation criteria ([Section 3.2](#)) and to conduct a meta-evaluation with our proposed framework ([Section 3.4](#)), we need an appropriate dataset of stories. In this chapter, we describe the building process of HANNA, our corpus of Human ANnotated Narratives. We begin by reviewing corpora related to Automatic Story Generation and find that no existing dataset contains human annotations w.r.t. specific criteria of story quality ([Section 4.2.1](#)). We then review and select ASG systems ([Section 4.2.2](#)) in order to build our own corpus, HANNA, specifically designed for ASG evaluation. The first version of HANNA contains 1,056 stories produced by 11 systems and aligned on 96 prompts ([Section 4.3.1](#)). We perform an annotation experiment on HANNA V1, asking human raters to grade each story w.r.t. our six human criteria. We analyze the overall and system-level correlations between our human criteria, and confirm that they allow for a standardized and extensive human evaluation ([Section 4.3.3](#)). We also observe that large pre-trained language models, namely GPT-2, produce the best results for ASG ([Section 4.3.4](#)). Finally, we use Large Language Models to augment HANNA with 480 new stories ([Section 4.4.1](#)) and 150k+ annotations ([Section 4.4.2](#)), and observe that LLMs are rated more highly than humans when evaluated by two different LLMs ([Section 4.4.3](#)).

Chapter 5: Meta-Evaluation Benchmark of Automatic Story Generation. After we have described the methodology for our meta-evaluation of ASG ([Chapter 3](#)) and built a corpus specifically designed for that task ([Chapter 4](#)), this chapter details our meta-evaluation of

automatic evaluation measures for Automatic Story Generation. First, we present our meta-evaluation of non-LLM automatic measures ([Section 5.1.1](#)) on HANNA V1, and a more fine-grained analysis of measures using specific methods ([Section 5.1.2](#)). We mainly observe that specific measures for Automatic Story Evaluation are needed, and that commonly used measures such as BLEU are sub-optimal. We then show our analysis of LLM performance at ASE: we compare LLMs to human and non-LLM automatic evaluation methods ([Section 5.2.1](#)) and we examine the influence of the Eval-Prompt on LLM performance ([Section 5.2.2](#)). We find that LLMs are currently the best proxy for human evaluation of ASG and that, in our specific setting, providing detailed guidelines does not improve correlations between LLM and human ratings.

Chapter 6: Exploring LLM Explainability for ASE and ASG. Since we observe in [Section 5.2](#) that LLMs correlate better with human judgment than other automatic measures, we wish to ascertain the reliability of our findings. In this chapter, we assess to which extent the performance displayed by LLMs at Automatic Story Evaluation can be explained through different factors. We begin with a related work on model explainability ([Section 6.2](#)). We perform two main experiments: a three-part study on LLM-generated explanations ([Section 6.3](#)), and an analysis of pretraining data on LLM performance ([Section 6.4](#)). For our analysis on LLM explanations, (1) we compute contextual embeddings for each explanation and perform a clustering analysis that shows that LLM explanations are overall well-separated w.r.t. our criteria ([Section 6.3.1](#)); (2) we perform a keyword analysis that confirms the previous findings ([Section 6.3.2](#)), and (3) we conduct a user study on LLM explanations to assess the appropriateness of the explanations ([Section 6.3.3](#)). Notably, we find that LLMs struggle to explain their answers with substantiated claims. Next, in order to check for model contamination and to evaluate the influence of pretraining data on the ASG performance of LLMs, we use a detection method based on minimum token probabilities ([Section 6.4](#)). We find that the observed performance of LLMs in our experiments is not a result of model contamination, and that larger models tend to produce content that is more similar to that of existing books.

Chapter 7: Conclusion. Finally, we summarize our main contributions and findings ([Section 7.1](#)), and outline three main research perspectives ([Section 7.2](#)): designing specific Automatic Story Evaluation measures, further investigating LLM performance at ASG and ASE, and assessing and mitigating the impact of LLMs on society.

1.4 Publications

This thesis has led to the publication of two research articles:

1. [Chhun et al., 2022] (**Chapter 3, Section 4.3 and Section 5.1**)

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

<https://aclanthology.org/2022.coling-1.509/>

[Note: COLING was ranked A at the time of acceptance of this publication.]

2. [Chhun et al., 2024] (**Section 4.4, Section 5.2 and Chapter 6**)

Cyril Chhun, Fabian M. Suchanek, Chloé Clavel. 2024. Do Language Models Enjoy Their Own Stories? Prompting Large Language Models for Automatic Story Evaluation. *Transactions of the Association for Computational Linguistics*, volume 12, pages 1122–1142.

https://doi.org/10.1162/tac1_a_00689

We also gave oral presentations for the NoRDF project¹ and for the “Journée Recherche TAL Défense” at the DGA Maîtrise de l’information Center.

1. **Cyril Chhun**. April 19th, 2023. Of Human Criteria and Automatic Metrics: A Benchmark of the Evaluation of Story Generation. *NoRDF Academic Workshop*, Télécom Paris, Palaiseau, France. Oral presentation.
2. **Cyril Chhun**. November 13th, 2023. Évaluation des histoires générées automatiquement. *Journée Recherche TAL Défense*, DGA Maîtrise de l’information, Rennes, France. Oral presentation.

¹<https://nordf.telecom-paris.fr/>

Chapter 2

Background

“ Long ago, the world was nothing more than an endless sea cloaked in a boundless sky, reaching as far as could possibly be imagined... ”

Shulk (*Xenoblade Chronicles*)

Contents

2.1 Pre-neural Automatic Story Generation	10
2.1.1 Structural Models	10
2.1.2 Planning-Based Models	12
2.1.3 Machine Learning-based Models	14
2.2 Language Models	14
2.2.1 Statistical Language Models	14
2.2.2 Neural Language Models	17
2.2.3 The Transformer Era	27
2.2.4 Pre-trained Language Models	29
2.2.5 Large Language Models	31
2.3 Evaluation and Meta-Evaluation	32
2.3.1 Human Evaluation	32
2.3.2 Automatic Evaluation	33
2.3.3 Meta-Evaluation	34
2.4 Conclusion	34

Chapter 2 Abstract

In this chapter, we provide some background information related to Automatic Story Generation and Evaluation (ASG and ASE). In [Section 2.1](#), we cover older methods for ASG: namely, structural and planning-based models. In [Section 2.2](#), we present language models, focusing on neural-based language models, from recurrent neural networks to large language models. Finally, in [Section 2.3](#), we present different aspects related to the question of evaluation in NLP.

2.1 Pre-neural Automatic Story Generation

In this section, we provide a brief history of previous work on automatic story generation which did not rely on neural networks; more detailed accounts can be found in [Alhussain and Azmi \[2021\]](#); [Kybartas and Bidarra \[2017\]](#). We re-use the classification of [Alhussain and Azmi \[2021\]](#), who identify three categories: structural models, planning-based models, and machine learning (ML) models.

2.1.1 Structural Models

A *story* can be defined as “a description of real or imaginary actors and events generated to achieve one or more goals, such as entertainment or education” [\[Alhussain and Azmi, 2021\]](#). In early research, a story was commonly viewed as a *script*, *i.e.*, a structure that describes a sequence of events that is coherent within a given context [\[Schank and Abelson, 1977\]](#). Such sequences, also called *schemas*, could then be used as building blocks for automatic story generation: after building a library of annotated stories, and given a schema divided into several slots, one can generate a story by retrieving matching slots from the library, as shown on [Figure 2.1](#).

One of the oldest structural models is found in [Propp \[1968\]](#). They analyzed around 600 Russian folktales and distilled from them a set of 31 character actions called *functions* (*e.g.* departure, struggle, rescue). Their work was used in Joseph Grimes’ pioneer system [\[Ryan,](#)

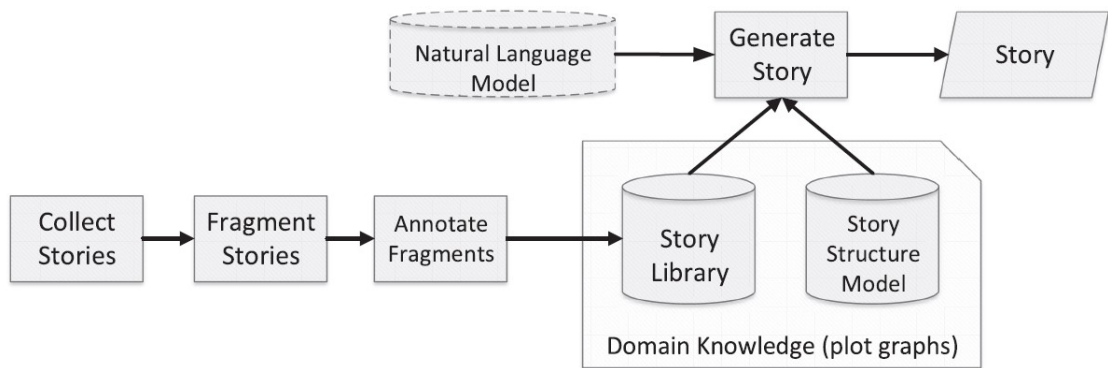


Figure 2.1: Structural story generation. Source: [Alhussain and Azmi \[2021\]](#).

A lion has been in trouble for a long time. A dog steals something that belongs to the lion. The hero, lion, kills the villain, dog, without a fight. The hero, lion, thus is able to get his possession back.

Table 2.1: The only surviving output from Joseph Grimes' system [Ryan, 2017].

2017], which selected random functions from Propp [1968] and ordered them according to the model's rules.

Scripts and schemas have been used in order to build **story graphs**. Maranda [1985] constructed a graph based on the model from Propp [1968] for the generation of folktales. The graph consists of nodes containing Propp's functions and of an annotated knowledge base that is searched for finding a good candidate for the next node. When a candidate is found, it is appended to the story, and this process is repeated until it reaches an end node.

Other works on narratives relied on theoretical structures called **story grammars**, which typically attempted to formalize how a sample of discourse is perceived as a coherent story, as opposed to another arrangement of the same sentences [Brewer and Lichtenstein, 1980]. While multiple story grammars have been introduced in the literature, they all postulate the necessary components of a story and the rules that describe their relations with one another.

Lakoff [1972] is the first to have used a grammar for story generation: they adapted Propp [1968] into a story grammar, viewing stories as words and Propp's functions as letters. Grasbon and Braun [2001] also relied on Propp's morphology but they did not completely automate plot generation, opting instead for guiding the author along a certain story path. Champagnat et al. [2010] used Campbell's Hero's Journey to create a generalized plot grammar capable of producing a high-level plot structure, also leaving the actual writing of the story to a human author. Colby [1973] developed a grammar focused on aboriginal folktales. It expanded the notion of events from Propp's model, which allowed for more unique and varied plot structures. Other grammars were proposed by Rumelhart [1975], Mandler and Johnson [1977], and Thorndyke [1977].

GESTER [Pemberton, 1989] is one of the earliest full-fledged grammar-based plot generators. Its grammar was derived from Old French epic narratives, describing both events with causes and effects but also the relations between characters. Bringsjord and Ferrucci [1999] proposed BRUTUS, which relied on plot grammars for the construction of betrayal narratives. Its stories could display high complexity but were specifically geared towards plots involving events of betrayal. Machado et al. [2001] developed TEATRIX, an interactive narrative system which created folktale templates based on Propp's functions. Author and interactor share the responsibility of the plot structure, assisted by the automatic selection of the next template and the automatic simulation of character actions. Spierling et al. [2002] introduced GEIST, which also creates templates for interactive stories using Propp's grammar. The stories are anchored in real world locations through augmented reality. SQUEGE [Onuczko et al., 2008] uses rewrite rules to progressively build more complex structures and selecting characters to fill plot roles in a manually authored space. Kybartas and Verbrugge [2014] combine grammar and graph-rewriting rules in their REGEN model, while also accounting for the relationships

between characters.

Structural models for story generation had the advantage of being straightforward and easy to implement, but they focused primarily on the correctness of the generated story's syntax, leaving aside its semantic aspect. Black and Wilensky [1979] argued that story grammars had major deficiencies which led them to conclude that they were not a promising approach to investigating story understanding. They suggested to explore more content-oriented approaches, e.g. planning knowledge.

2.1.2 Planning-Based Models

Since story grammars did not account for the semantic relationships within a story, their scope could not extend to stories with multiple protagonists or conflicting goals. Moreover, they sometimes acknowledged non-stories as correct stories [Black and Bower, 1980]. To tackle those issues, Wilensky [1983] proposed the story points theory, which takes into account story semantics. Here, the story is seen as a chain of causally connected events to pursue an end goal. This led to the development of planning-based models for story generation.

■ Goal-Directed Approaches

The first category of planning-based models adopted goal-directed approaches. Meehan [1977] introduced TALE-SPIN, a story generation program that simulates rational behaviour by characters in the fictional world. It used a problem solver that, given a goal, was able to generate subgoals and actual events in a repeated fashion until the character goal is reached. It could generate short and coherent stories, but those stories often lacked climax or resolution.

A solution to this issue was to move the focus from character goals to *global goals* directly imagined by the author, as it should lead to an improved story structure. Since authors are exterior to the story world, their intention to produce a good story can result in goal conflicts between characters, increasing narrative tension. AUTHOR [Dehn, 1981] is one of the first story generation models based on author goals. It aimed at simulating a human author: using a reconstructive dynamic memory architecture, AUTHOR wrote a story draft based on a set of author goals and revised it multiple times before delivering it. As a result, it produced stories with better structures, but its characters were less believable since they sometimes acted without clear intentions to satisfy the author's goals.

Thus, *multi-agent approaches* were considered, where story generation is directed both by character and author goals. This combination of different levels of planning has been achieved with several mechanisms. Theune et al. [2003] used environmental and motivational control (e.g. by introducing new character or character goals) preserving the story's simple structure. Riedl and Young [2004] proposed an Intent-driven Partial Order Causal-Link (IPOCL): given a set of author goals, it infers character actions through causal links. It then estimates the believability of those actions for evaluating the quality of the overall story plan, at the cost of a long generation time [Riedl and Young, 2010]. Haslum [2012] modeled character intentions as preconditions of character actions, allowing for the use of pre-existing planners to generate stories. Ware and Young [2011] improved IPOCL by creating a model of conflict and enforcing the generation of conflict in stories by constraining the planner.

■ Analogy-Based Approaches

Computational analogy operates by identifying similarities and transferring knowledge between a source domain and a target domain [Zhu and Ontanón, 2010]. Analogies can be used at different levels: events, characters, stories, discourse, narratives.

Turner [1993] introduce MINSTREL, a multi-agent system that uses Case-Base Reasoning (CBR). MINSTREL stores scenes in an episodic memory and indexes them by location and action to form groups. At generation time, it first searches for a matching scene in memory; if there is none, it creates a new one.

MEXICA [Pérez and Sharples, 2001] creates a story world context to record emotional relationships between characters and the main dramatic tensions of the story. Those elements are then used to condition the story actions, as each schema of elements is associated with a possible set of actions. During generation, the memory is searched for a matching schema; if one is found, an associated action is selected and the story is updated accordingly.

Gervás et al. [2005] reuse Propp's story morphology to design their CBR-based system PROTOPROPP. They modeled an ontology of Propp's functions that describes their temporal relationships and co-occurrence constraints and made the generation process interactive where the user provides inputs (e.g. functions and attributes).

■ Heuristic Search Approaches

Since planning-based approaches used diverse constraints for generating stories, their search domain had to be restricted for the sake of efficiency, resulting in stories that could lack variety. Heuristic Search Approaches tackled that issue by expanding the search space.

Ong and Leggett [2004] proposed HEFTI, which uses genetic algorithms for story generation. It represents story components as chromosomes, where story components are defined as sets of story elements that are linked together, scripts that control the characters, and variables that record the state of the story. Each chromosome is composed of genes, which are subsets of story elements. Crossovers only occur at the chromosome level, while mutations occur at both chromosome and gene levels.

McIntyre and Lapata [2010] also use genetic algorithms. They extracted plot graphs from a corpus of stories where nodes represent events. An example is shown on Figure 2.2. Given a sentence, an initial population of story plots is sampled from the corresponding plot graph; those plots are represented as an ordered graph of dependency trees. Then, crossovers and mutations are applied on the plot trees. The fitness of the plots is based on the coherence of the story.

Kartal et al. [2014] formulated story generation as a Monte Carlo tree search problem. Story states are represented as nodes in the tree, while edges represent actions, which make the story move from state to state. An evaluation function considers two factors: the percentage of user goals achieved in the story and the product of the believabilities of each action. The best story is then selected.

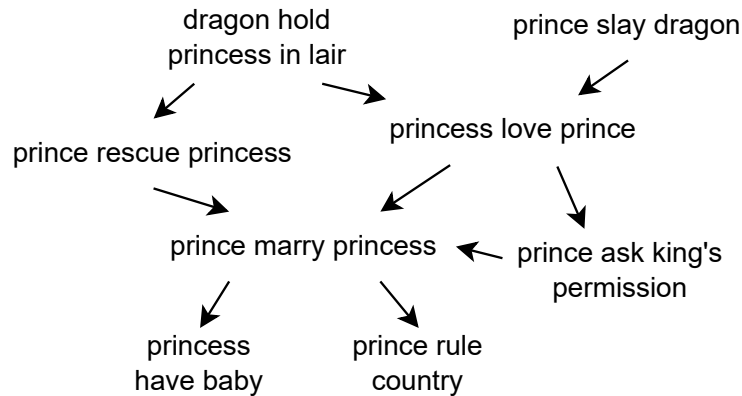


Figure 2.2: Plot graph for the sentence: “the princess loves the prince” [McIntyre and Lapata, 2010].

2.1.3 Machine Learning-based Models

Contrary to the rule-based ASG systems presented in [Section 2.1.1](#) and [Section 2.1.2](#), machine learning-based models rely on probabilistic strategies. In recent years, the vast majority of machine learning-based models have been employing language models. In the next section, we provide an overview of the evolution of language models. We review the models specific to ASG later, in [Section 4.2.2](#).

2.2 Language Models

Language models are probabilistic models of natural language. More precisely, their function is to estimate the probability distribution of the next word for a given sequence of words. The first developed language models were called statistical language models, for they relied on statistical estimation techniques applied to textual data. In the 2010s, language models relying on neural architectures started to achieve state-of-the-art performance in many Natural Language Processing (NLP) tasks (*e.g.* machine translation, named entity recognition, semantic role labeling, question answering, etc.), and they have now become omnipresent in NLP research topics, including ASG. In this section, we will describe the most noteworthy advances in the development of language models, from statistical language models to the current state-of-the-art models, which are called Large Language Models (LLMs).

2.2.1 Statistical Language Models

■ Definitions

Statistical Language Models (SLM) attempt to capture regularities of natural language for the purpose of improving the performance of various natural language applications [Rosenfeld, 2000]. Before the advent of neural models, they were commonly used for machine translation, document classification, information retrieval, and many other tasks.

A SLM is, at its core, a probability distribution $P(s)$ over all possible sentences s . SLMs were usually used in the context of a Bayes' classifier, where they could play the role of either the prior or the likelihood function. For instance, in automatic speech recognition, given an acoustic signal a , the goal is to find the sentence s^* that is most likely to have been spoken. With a Bayesian framework, the solution is:

$$s^* = \arg\max_s P(s | a) = \arg\max_s (P(a | s) \cdot P(s))$$

where the language model $P(s)$ plays the role of the *prior*.

By contrast, in document classification, given a document d , the goal is to find the most probable class c^* . Typically, for each class c , we have a corresponding language model $P_c(d)$. The solution is then:

$$c^* = \arg\max_c P(c | d) = \arg\max_c (P(d | c) \cdot P(c)) = \arg\max_c (P_c(d) \cdot P(c))$$

where the language models $P_c(d)$ play the role of the *likelihood*.

The quality of a given SLM is commonly assessed using the *cross-entropy* of a new data sample.

Definition 1: Cross-entropy (SLM)

Let P be a statistical language model and $X = \{w_1, \dots, w_N\}$ a sentence of N words. The cross-entropy of X is defined by:

$$\text{CE}(X) := \frac{1}{N} \sum_i -\log_2 P(w_i).$$

Cross-entropy can be interpreted as the average number of bits that would be required to encode the sample data using an optimal coder. Assuming the data is generated by a random process, a perfect model would have the lowest cross-entropy possible. Thus, we generally aim to decrease model cross-entropy, as it is used as an inverse proxy for model quality.

In practice, the performance of a SLM is often reported in terms of *perplexity*.

Definition 2: Perplexity (SLM)

Let P be a statistical language model and $X = \{w_1, \dots, w_N\}$ a sentence of N words. The perplexity of X is defined by:

$$\text{PPL}(X) := 2^{\text{CE}(X)} = 2^{\frac{1}{N} \sum_i -\log_2 P(x_i)}.$$

Perplexity can be interpreted as the average branching factor of the language according to the model, *i.e.*, the weighted average number of choices for each word. As perplexity becomes lower, the model tends to make more accurate predictions, although there have been occurrences in the literature of negative correlations between perplexity and error rate [Chen et al., 1998].

A variety of SLM techniques has been developed, such as n -gram models, decision tree models, context-free grammars, exponential models, adaptive models, etc. We will here only describe n -gram models, which were by far the most popular forms of SLM. A more detailed account can be found in [Rosenfeld \[2000\]](#).

■ n -gram Models

For a sentence s composed of words $w_1, \dots, w_l$, we can express its probability $P(s)$ by:

$$\begin{aligned} P(s) &= P(w_1)P(w_2 | w_1)P(w_3 | w_1, w_2) \dots P(w_l | w_1, \dots, w_{l-1}) \\ &= P(w_1) \prod_{i=2}^l P(w_i | w_1, \dots, w_{i-1}). \end{aligned}$$

The intuition of the n -gram model is that instead of computing the probability of a word given all its predecessors, we assume that the probability depends only on the $n - 1$ previous words. For instance, for $n = 3$, we assume that:

$$P(w_i | w_1, \dots, w_{i-1}) \approx P(w_i | w_{i-2}, w_{i-1}).$$

The value of n trades off the variance of the estimate against its bias. Incidentally, the trigram assumption ($n = 3$) has been shown to work well in practice [[Goodman, 2001](#)] and is a common choice with large training corpora, while bigrams ($n = 2$) are used with smaller ones. Trigram probabilities $C(w_{i-2}, w_{i-1}, w_i)$ are usually estimated from counts on a training corpus, and similarly for bigram probabilities $C(w_{i-2}, w_{i-1})$. Then, we can approximate:

$$P(w_i | w_{i-2}, w_{i-1}) \approx \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})}.$$

A glaring weakness of n -gram models is that this approximation is very noisy, since many n -grams do not appear in the training data. The probability of unseen n -grams therefore defaults to 0, which can severely impede model performance. To mitigate this phenomenon, smoothing strategies were introduced [[Chen and Rosenfeld, 2000](#)]. For instance, Kneser-Ney smoothing [[Kneser and Ney, 1995](#)], commonly seen as the overall best smoothing strategy, accounts for unseen n -grams by subtracting a fixed number D from all n -gram counts. It then estimates the probability of an unseen n -gram by interpolating the probabilities of all lower-order n -grams: if any k -gram of a sequence was seen in the corpus, the sequence will have a non-zero probability.

Several other techniques have been developed to improve n -gram models, such as skipping, clustering, caching, sentence mixture, etc. [[Goodman, 2001](#)].

A few attempts to use n -gram models for story generation were made, with limited success. [Swanson and Gordon \[2009\]](#) describe a system that takes an input sentence from a user and searches in a database for a sentence that continues the story. Specifically, the retrieval procedure uses a context bigram model. [Limpanadusadee et al. \[2014\]](#) evaluate the performance of n -gram models for storytelling sentence generation in Thai: 3- or 4- word sentence generation yielded between 40% and 100% accuracy, dropping to as low as 10% for 6-word sentences.

2.2.2 Neural Language Models

■ A Brief History of Neural Networks

Neural networks are algorithms used for machine learning, which is defined as the field of study that gives computers the ability to learn without being explicitly programmed [Samuel, 1959]. Neural networks take inspiration from the human brain’s architecture, where a neuron’s computations involve a weighted sum of its input values, to which a nonlinear function is applied. The first neural network model was proposed by McCulloch and Pitts [1943], who proved that any finite logical expression can be realized by a network of artificial neurons. Then, Hebb [1949] coined the term “connectionism” and described a method of updating synaptic weights, as well as other contributions which are now implemented in neural network models, at least to some degree. Rosenblatt [1958] defined a neural network structure called the perceptron, which is considered a landmark in the history of neural networks and laid the groundwork for both supervised and unsupervised learning. Widrow and Hoff [1960] introduced ADALINE, a neural network model with a different learning algorithm from the perceptron, of which an extension is used in today’s backpropagation networks. Neural networks were then forgotten for a few decades, in part due to a campaign that discredited their potential [Minsky and Papert, 1969]. In the 1980s, several events sparked a regain of interest in neural networks: Hopfield [1982] presented recurrent artificial neural networks serving as content-addressable memory systems, Rumelhart et al. [1988] rediscovered the backpropagation algorithm (first discovered by Werbos [1974]) and Le Cun et al. [1990] proposed the LeNet network for hand-written digit recognition. As neural networks became deeper, the backpropagation algorithm became slower and had a glaring weakness: it suffered from the vanishing gradient problem. This was eventually solved when multiple studies (e.g. Nair and Hinton [2010], Glorot et al. [2011]) advised to use the rectified linear unit (ReLU) activation function. The early 2010s then saw a significant rise in usage of neural networks, helped by a growing abundance of training data and an increased availability of both computational power and graphics processing units.

■ Basic Notions

Definition 3: Neural Network

A neural network is a machine learning architecture which maps a vectorized input $\mathbf{x}$ to a vectorized output $\mathbf{y}$ through a parameterized function f :

$$\hat{\mathbf{y}} = f(\mathbf{x}, \theta),$$

where $\hat{\mathbf{y}}$ is the predicted output produced by the function f with parameters θ .

The building block of neural networks is the *neuron*, whose inputs x_i are weighted by learned parameters w_i and linearly combined with a bias b before the application of a non-linear function. The first neuron algorithm to be implemented on real hardware was the perceptron, proposed by Rosenblatt [1958].

Definition 4: Perceptron

The perceptron is an algorithm for learning a binary classifier called a threshold function: it maps its input $\mathbf{x}$ (a real-valued vector) to an output value $f(\mathbf{x})$ (a single binary value):

$$f(\mathbf{x}) = \theta(\mathbf{w} \cdot \mathbf{x} + \mathbf{b}),$$

where θ is the Heaviside step-function, $\mathbf{w}$ a vector of real-valued weights and $\mathbf{b}$ a bias vector.

Following the introduction of the perceptron, different types of architectures for neural networks have been proposed, such as the *feed-forward neural network*, the *convolutional neural network* (CNN) and the *recurrent neural network* (RNN).

Feed-Forward Neural Network. The most basic type of neural network is the feed-forward neural network, where neurons are arranged in multiple layers. Each neuron receives as inputs the outputs from the neurons of the previous layer. For a *fully-connected* feed-forward neural network containing two hidden layers, the function can be summarized as:

$$\hat{\mathbf{y}} = \sigma(\mathbf{W}_3 \mathbf{h}_2 + \mathbf{b}_3)$$

$$\mathbf{h}_2 = f_2(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2)$$

$$\mathbf{h}_1 = f_1(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1),$$

where

- $\mathbf{x}$ is the input vector,
- $\mathbf{h}_1$ and $\mathbf{h}_2$ are the first and second layers respectively,
- $\hat{\mathbf{y}}$ is the output vector,
- f_1 and f_2 are non-linear functions (*e.g.* the sigmoid function, the hyperbolic tangent or the rectified linear unit (ReLU))
- the $\mathbf{W}_i$ and $\mathbf{b}_i$ for $i \in \{1, 2, 3\}$ are the learned weights,
- σ is the softmax function defined as:

$$\sigma: \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$\mathbf{z} \mapsto \left(\frac{e^{(z_j)}}{\sum_{k=1}^K e^{(z_k)}} \right)_{j \in \{1, \dots, n\}}.$$

A corresponding visual example can be seen on [Figure 2.3](#).

Convolutional Neural Networks (CNN). CNNs, like feed-forward neural networks, consist of an input layer, hidden layers and an output layer. In CNNs, the hidden layers include one or more layers that perform convolutions, *i.e.*, layers that perform a dot product of the

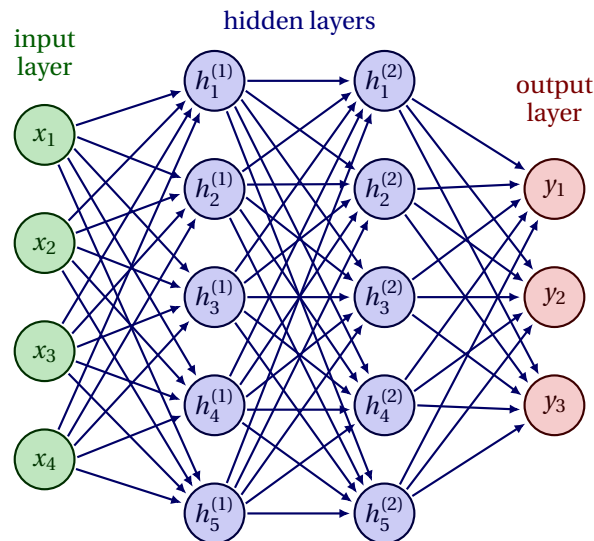


Figure 2.3: A fully-connected feed-forward neural network [Neutelings, 2021].

convolution kernel with the layer's input matrix. As the convolution kernel slides along the input matrix for the layer, the convolution operation generates a feature map, which in turn contributes to the input of the next layer, as shown on Figure 2.4. Usually, the convolutional layers are followed by other layers, *e.g.* pooling layers, fully-connected layers, and normalization layers. While CNNs have sometimes been used for natural language processing tasks, they are mainly used in computer vision settings, *e.g.* for image classification or object recognition.

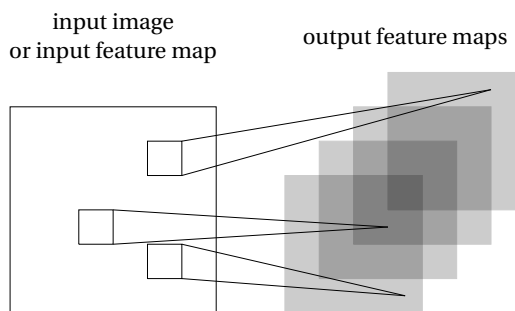


Figure 2.4: A layer of a convolutional neural network [Stutz, 2020].

Recurrent Neural Networks (RNN). RNNs are bi-directional neural networks: the output from some nodes can affect subsequent input to the same nodes. RNNs can use internal states to process arbitrary sequences of inputs, which makes them well-suited for tasks related to natural language processing. RNNs typically share parameters across each layer of the network and leverage backpropagation through time for their training procedure.

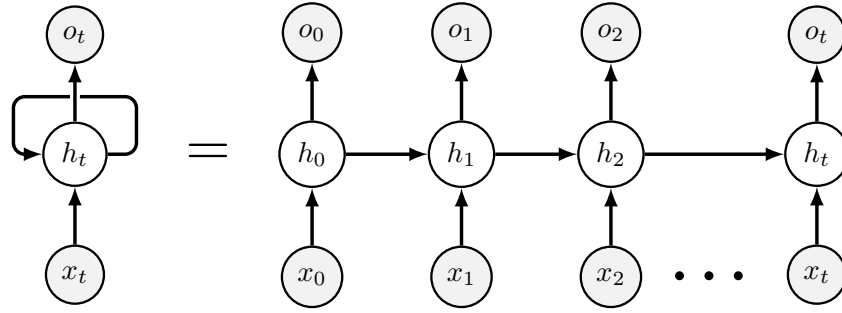


Figure 2.5: A fully-recurrent neural network [User121799, 2019].

A basic fully-recurrent neural network (shown on Figure 2.5) can be described by the two following formulas:

$$\begin{aligned}\mathbf{h}_{t+1} &= f(\mathbf{W}_x \cdot \mathbf{x}_t + \mathbf{W}_h \cdot \mathbf{h}_t + \mathbf{b}_h) \\ \mathbf{o}_{t+1} &= f(\mathbf{W}_o \cdot \mathbf{h}_t + \mathbf{b}_o),\end{aligned}$$

where

- $\mathbf{x}_t$ is the input vector at time t ,
- $\mathbf{h}_t$ is the hidden state vector at time t ,
- $\mathbf{o}_t$ is the output vector at time t ,
- the $\mathbf{W}$'s and $\mathbf{b}$'s are weight matrices and bias vectors parameters learned during training and shared across time steps.

However, the high depth of RNNs can make them subject to the vanishing gradient problem, where the gradients computed for updating the weights would have near-zero values. To solve this issue, different architectures were proposed, such as the *long short-term memory* (LSTM) network.

Long Short-Term Memory (LSTM). An LSTM network [Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2012] is composed of multiple units, with each unit containing a cell, an input gate, an output gate and a forget gate. The three gates control the information flow into and out of the cell, and the cell retains values for arbitrarily long periods of time. Forget gates use a coefficient between 0 and 1 to decide at which rate information from a prior state should be discarded w.r.t. the present input. Using the same mechanism as forget gates, input gates determine which new pieces of information to store in the existing state, and output gates regulate the information to output. This architecture enables LSTM networks to retain valuable long-term dependencies.

Formally, an LSTM network (see Figure 2.6) computes a mapping from an input sequence $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ to an output sequence $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_T)$ using the following equations iteratively

from $t = 1$ to T :

$$\begin{aligned}
 \mathbf{f}_t &= \sigma_g(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f) \\
 \mathbf{i}_t &= \sigma_g(\mathbf{W}_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i) \\
 \mathbf{o}_t &= \sigma_g(\mathbf{W}_o \mathbf{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o) \\
 \tilde{\mathbf{c}}_t &= \tanh(\mathbf{W}_c \mathbf{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \\
 \mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \\
 \mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \\
 \mathbf{y}_t &= \mathbf{h}_t
 \end{aligned}$$

where

- $\mathbf{x}_t$ is the input vector,
- $\mathbf{f}_t$ is the forget gate's activation vector,
- $\mathbf{i}_t$ is the input gate's activation vector,
- $\mathbf{o}_t$ is the output gate's activation vector,
- $\mathbf{h}_t$ is the hidden state vector,
- $\mathbf{y}_t$ is the output vector,
- $\tilde{\mathbf{c}}_t$ is the cell input activation vector,
- $\mathbf{c}_t$ is the cell state vector,
- the $\mathbf{W}$'s, $\mathbf{U}$'s and $\mathbf{b}$'s are weight matrices and bias vectors parameters learned during training,
- σ_g is the sigmoid function.

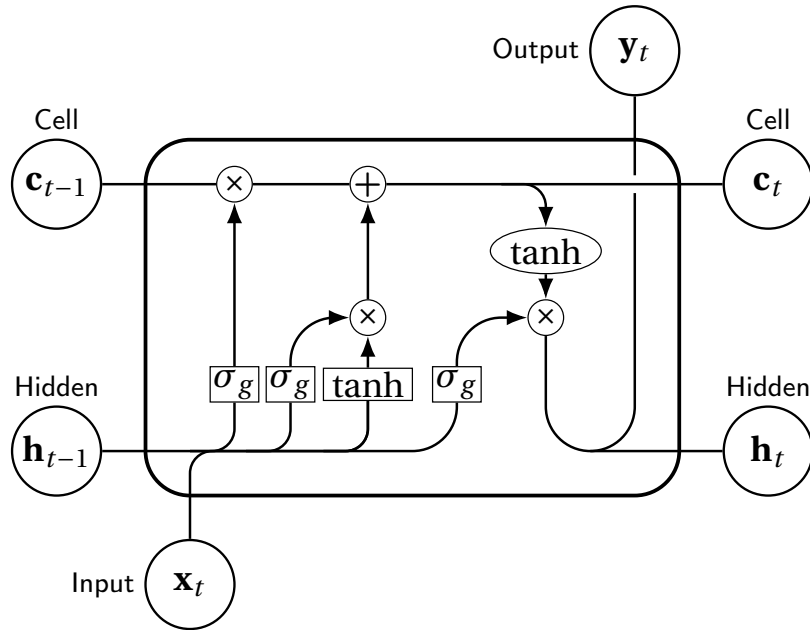


Figure 2.6: An LSTM cell [V, 2018].

■ Neural Network Training

In order to learn the best parameters, one needs to *train* a neural network. It is typically done through the minimization of a loss function over a training set, typically with a gradient-based method [Goldberg, 2016]. Most such methods repeatedly compute an estimate of the error over the dataset, then compute the gradient with respect to the error, and update the parameters by moving them in the direction opposite to the gradient.

Loss Functions. A loss function L assigns a numerical score $L(\hat{\mathbf{y}}, \mathbf{y})$ which estimates the error between the network's output $\hat{\mathbf{y}}$ and the true expected output $\mathbf{y}$. The loss function should be bounded from below, with the minimum attained only for cases where the network's output is the true expected output.

In regression settings, a common loss is the *mean squared error* (MSE), defined by:

$$\text{MSE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i, y_i)^2.$$

The variability of the gradient of the MSE loss is beneficial to the fast convergence and high accuracy of the model, but makes it very sensitive to outliers.

For classification settings, which are more common in NLP, the categorical cross-entropy loss (or negative log-likelihood) is typically used. It is defined by:

$$L_{\text{cross-entropy}}(\hat{\mathbf{y}}, \mathbf{y}) = - \sum_i y_i \log(\hat{y}_i),$$

where $\mathbf{y}$ is the true multinomial distribution over the labels $1, \dots, n$ and $\hat{\mathbf{y}}$ is the network's output after the softmax activation function, *i.e.*, the class membership conditional distribution: $\hat{y}_i = P(y = i \mid \mathbf{x})$.

The output of this loss provides a probabilistic interpretation and the softmax function makes it quite stable, but it can be sensitive to imbalanced datasets, where the number of samples for each class is not equal.

We refer the reader to Wang et al. [2022] for a more comprehensive discussion on loss functions in machine learning.

Optimization Algorithms. Early attempts at minimizing the loss function (*e.g.* [Rumelhart et al., 1988]) used the gradient descent algorithm, in which the weights are iteratively updated according to the following formula:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \gamma \nabla L(\hat{\mathbf{y}}, \mathbf{y}),$$

where γ is an adequately chosen parameter. Under sufficient regularity assumptions, when the initial estimate $\mathbf{W}_0$ is close enough to the optimum, and when γ is sufficiently small, the gradient descent achieves linear convergence [Dennis Jr. and Schnabel, 1996].

Since then, numerous other approaches have been proposed: stochastic gradient descent, batch gradient descent, ADAGRAD [Duchi et al., 2010], RMSPROP [Tieleman and Hinton, 2012], ADAM [Kingma and Ba, 2015], etc.

■ First Neural Models

The goal of statistical language modeling is to learn the joint probability function of sequences of words in a language, where words are modeled by discrete random variables. Bengio et al. [2000] observed that this objective is intrinsically difficult because of the curse of dimensionality: it is likely that word sequences encountered by the model will not have been seen during training. They proposed a recurrent neural model that learned a distributed representation for words along with the probability distribution of word sequences: in particular, the model associates with each word in the vocabulary a *feature vector*, i.e., a real-valued vector in $\mathbb{R}^m$ where m is a chosen size. The idea was further refined by Mikolov et al. [2010], and showed significant improvements in performance compared to n -gram models in terms of perplexity. While feed-forward neural networks had also been tested for language modeling [Xu and Rudnicky, 2000], RNN-based models had an inherent advantage in their ability to process sequence data of variable lengths, and the sharing of parameters across layers reduced the number of computations compared to feed-forward neural networks [Jing and Xu, 2019].

■ Static Word Embeddings

Mikolov et al. [2013a] proposed the WORD2VEC architecture for computing continuous vector representations of words. Contrary to previous count-based approaches such as one-hot encoding, the WORD2VEC approach allows for language models to learn high-quality word vectors from datasets containing billions of words, with millions of words in the vocabulary. In particular, it was shown that the computed representations of words displayed multiple degrees of similarity. Two variations were proposed: continuous bag of words (CBOW) and Skip-gram (see Figure 2.7). The main motivation behind those new architectures was to reduce computational complexity: a simpler model would be trained to produce word representations that could then be used by more complex RNN-based models. In CBOW, the context word vectors are projected into the same position in order to produce an averaged vector, which is used to predict the target word. In Skip-gram, the task is reversed: the model is shown the target word and has to predict the context words. They show that the learned word vectors from their models outperform the previous state-of-the-art on relational similarity tasks.

Mikolov et al. [2013b] improved WORD2VEC through the use of multiple strategies: hierarchical softmax (which allows for evaluation of only $\log_2(W)$ output nodes instead of W nodes for obtaining the probability distribution), subsampling of frequent words (deleting frequent words from the training corpus), and negative sampling (updating only a limited number of incorrect words for each training instance).

Other static word embeddings built upon the WORD2VEC approach. GLOVE embeddings [Pennington et al., 2014] combine the advantages of the two major model families in the literature: global matrix factorization and local context window methods. They leverage statistical information by training only on the nonzero elements in a word-word co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus. FASTTEXT embeddings [Bojanowski et al., 2017] are based on the Skip-gram model, where each word is represented as a bag of character n -grams. A vector representation is associated to each character n -gram and words are represented as the sum of these representations.

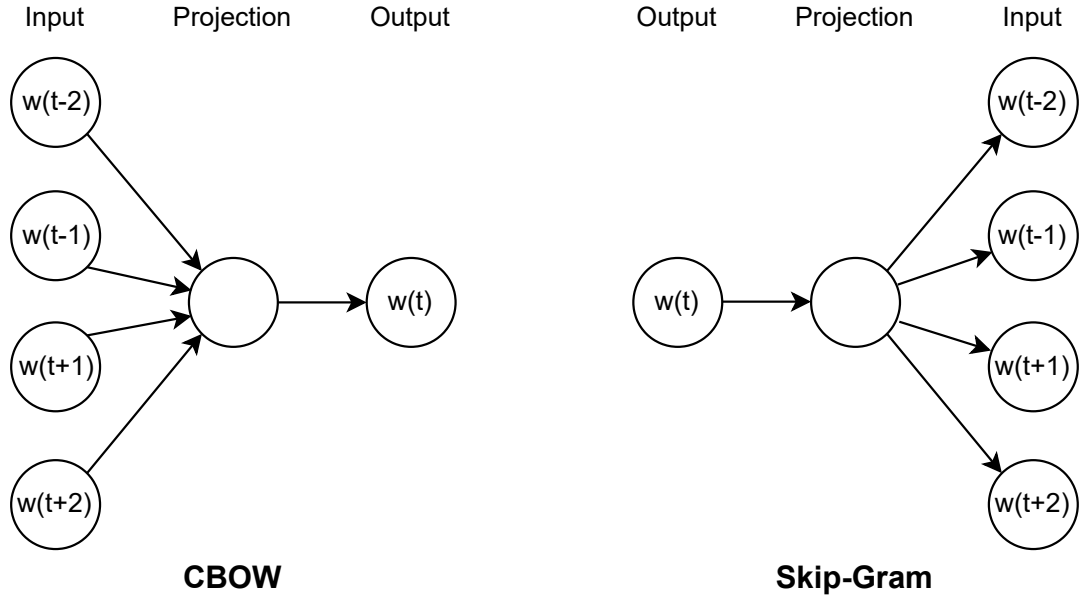


Figure 2.7: CBOW and Skip-gram architectures [Mikolov et al., 2013a].

■ RNN Encoder-Decoder

While RNNs are powerful and flexible, they could only be applied to tasks where inputs and outputs are of fixed lengths. However, many NLP tasks, such as machine translation, involve variable-length inputs and outputs. To solve this issue, [Choi et al. \[2014\]](#) and [Sutskever et al. \[2014\]](#) introduced the *Encoder-Decoder* framework, also called *Sequence-to-sequence* (SEQ2SEQ) framework. This architecture consists of two RNNs: the first RNN encodes a sequence of symbols into a fixed-length vector representation, and the second RNN decodes that representation into another sequence of symbols.

Formally, the Encoder-Decoder can be described as follows: an encoder reads the input sentence $\mathbf{x} = (x_1, \dots, x_T)$ into a vector c . Commonly, a RNN is used such that

$$h_t = f(x_t, h_{t-1}) \quad \text{and} \quad c = q(h_1, \dots, h_T),$$

where h_t is the hidden state at time t , and f and q are nonlinear functions. For instance, [Sutskever et al. \[2014\]](#) used an LSTM as f and $q(h_1, \dots, h_T) := h_T$.

Then, the decoder is trained to predict the next word $y_{t'}$ given the context vector c and the previously predicted words $y_1, \dots, y_{t'-1}$, *i.e.*, it defines a probability over the output sequence $\mathbf{y}$ by decomposing the joint probability into the ordered conditionals:

$$p(y_1, \dots, y_{T_y} \mid x_1, \dots, x_{T_x}) = \prod_{t=1}^{T_y} p(y_t \mid c, y_1, \dots, y_{t-1}).$$

With an RNN, each conditional probability is modeled as

$$\prod_{t=1}^{T_y} p(y_t | c, y_1, \dots, y_{t-1}) = g(y_{t-1}, s_t, c),$$

where g is a nonlinear function that outputs the probability of y_t , and s_t is the hidden state of the RNN.

In particular, [Sutskever et al. \[2014\]](#) showed that a large deep LSTM Encoder-Decoder model with limited vocabulary and few assumptions about problem structure was able to outperform statistical language models for machine translation tasks.

■ Attention Mechanism

[Bahdanau et al. \[2015\]](#) conjectured that the Seq2seq framework's encoding of the input sequence into a fixed-length vector was a bottleneck for performance improvement, since it may limit its ability to process long sentences [\[Cho et al., 2014\]](#). Thus, they extended the architecture by allowing the model to automatically soft-search for parts of the input sequence that are relevant to predicting a target word, which would later be called the *attention mechanism*. The new model they propose for machine translation consists of a bidirectional RNN as an encoder and a decoder that emulates searching through a source sentence when decoding a translation.

The encoder is supposed to annotate each word by summarizing both the preceding and following words, hence they choose a bidirectional RNN. The forward RNN $\vec{f}$ reads the input sequence $(x_1, \dots, x_{T_x})$ from left to right and computes a sequence of forward hidden states $(\vec{h}_1, \dots, \vec{h}_{T_x})$. Conversely, the backward RNN $\overleftarrow{f}$ reads the input from right to left and computes backward hidden states $(\overleftarrow{h}_1, \dots, \overleftarrow{h}_{T_x})$. The annotation $\mathbf{h}_j$ for each word x_j is obtained by concatenation of both hidden states:

$$h_j = \left[\vec{h}_j^\top ; \overleftarrow{h}_j^\top \right]^\top.$$

In the decoder, each conditional probability for a word y_i is defined as

$$p(y_i | y_1, \dots, y_{i-1}, c_i) = g(y_{i-1}, s_i, c_i),$$

where g is a non-linear (potentially multi-layered) function and s_i is the RNN hidden state for time i computed by

$$s_i = f(s_{i-1}, y_{i-1}, c_i).$$

Unlike in the RNN Encoder-Decoder, there is a context vector c_i for each word y_i .

c_i is computed as a weighted sum of the annotations $(h_1, \dots, h_{T_x})$ produced by the encoder:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad \text{where} \quad \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad \text{where} \quad e_{ij} = a(s_{i-1}, h_j)$$

is an alignment model which scores how well the inputs around position j and the output at position i match. The score is based on the previous hidden state s_{i-1} and the j -th annotation h_j of the input.

[Bahdanau et al. \[2015\]](#) originally chose to parameterize a as a feedforward neural network jointly trained with the rest of the architecture:

$$a(s_{i-1}, h_j) = v_a^T \tanh(W_a[s_{i-1}; h_j]).$$

Other alignment scores have been proposed:

1. Content-based [[Graves et al., 2014](#)]: $a(s_{i-1}, h_j) = \cos(s_{i-1}, h_j)$;
2. Location-based [[Luong et al., 2015](#)]: $\alpha_{ij} = \text{softmax}(W_a s_{i-1})$ (no dependency on the source position);
3. General [[Luong et al., 2015](#)]: $a(s_{i-1}, h_j) = s_{i-1}^T W_a h_j$;
4. Dot-product [[Luong et al., 2015](#)]: $a(s_{i-1}, h_j) = s_{i-1}^T h_j$;

The context vector c_i can be understood as an expected annotation with α_{ij} being the probability that y_i is aligned with x_j . Meanwhile, α_{ij} reflects the importance of the annotation h_j with respect to the previous hidden state s_{i-1} in deciding the next s_i and generating the next word y_i , *i.e.*, it reflects how much the decoder should pay “attention” to the annotation.

■ Contextual Embeddings

[Peters et al. \[2018\]](#) introduce ELMo (Embeddings from Language Models), a new type of deep contextualized word representation that is able to model both (1) complex characteristics of word use (*e.g.* syntax and semantics), and (2) how these uses vary across linguistic contexts (*i.e.*, to model polysemy). They demonstrate that this representation can be easily integrated into existing models, and significantly improves the state of the art in every considered case across a range of challenging language understanding problems.

ELMo is designed as a task specific combination of the intermediate layer representations of a bidirectional language model (biLM). For each token t_k , a L -layer biLM computes a set of $2L + 1$ representations: the initial token embedding x_k , and the respectively forward and backward embeddings $\vec{h}_{k,j}$ and $\overleftarrow{h}_{k,j}$ for each layer j .

Let $h_{k,0} := x_k$ and $h_{k,j} := [\vec{h}_{k,j}; \overleftarrow{h}_{k,j}]$. ELMo is defined as a weighted average of all biLM layers:

$$\text{ELMo}_k := \gamma^{\text{task}} \sum_{j=0}^L s_j^{\text{task}} h_{k,j},$$

where the s_j^{task} are softmax-normalized weights and the scalar parameter γ^{task} allows the task model to scale the entire ELMo vector.

Through extensive experiments, they demonstrate that ELMo representations work extremely well in practice. They first show that they can be easily added to existing models for six diverse language understanding problems, such as question answering and sentiment analysis, and that the addition of ELMo representations alone significantly improves the state of the art in every case, including up to 20% relative error reductions.

2.2.3 The Transformer Era

The introduction of the transformer architecture in the landmark paper “Attention Is All You Need” [Vaswani et al., 2017] marked a notable shift in NLP, as transformers would become the foundation of all future advances in the development of language models. Departing from the usual RNN-based or rarer CNN-based models, they proposed an auto-regressive architecture that relied entirely on attention mechanisms. Notably, the transformer achieved new state of the art performance in the WMT 2014 tasks.

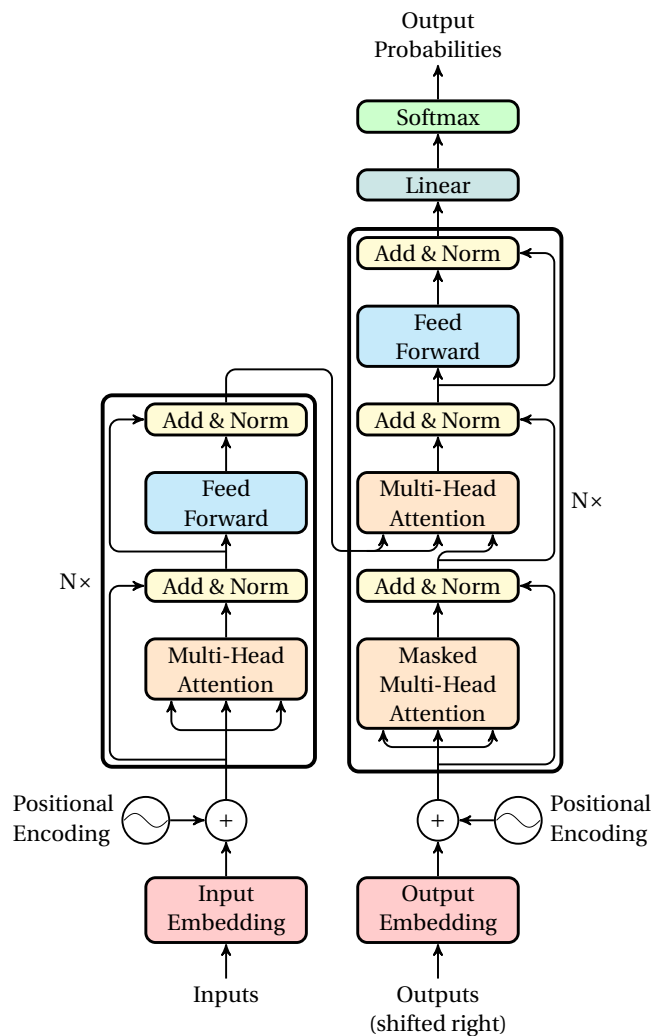


Figure 2.8: The transformer architecture [Vaswani et al., 2017].

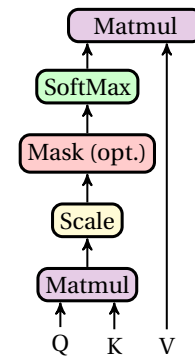


Figure 2.9: Scaled dot-product attention [Vaswani et al., 2017].

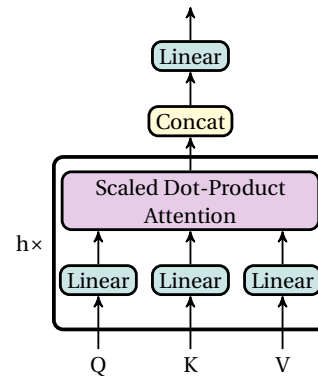


Figure 2.10: Multi-head attention [Vaswani et al., 2017].

■ Encoder and Decoder Stacks

The transformer, like the RNN encoder-decoder, is composed of an encoder and a decoder (see Figure 2.8 for an illustration). The encoder consists of N stacks of identical encoder layers containing two sub-layers: a multi-head self-attention mechanism and a feed-forward network. The decoder contains the same sub-layers and inserts a third sub-layer which performs multi-head attention over the output of the encoder stack. Also, this sub-layer is masked to prevent positions from attending to subsequent positions, *i.e.*, to ensure that predictions for position i depend only on positions less than i . Finally, residual connections are employed around each sub-layer, followed by layer normalization.

■ Scaled Dot-Product Attention

The transformer uses a particular attention mechanism called scaled dot-product attention (see Figure 2.9), where the input consists of queries and keys of dimension d_k and values of dimension d_v . Queries, keys, and values are packed together into matrices Q , K , and V , and the output is computed as follows:

$$\text{Attention}(Q, K, V) := \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

The main advantage of scaled dot-product attention over previous attention functions is that it is faster, more space-efficient, and can be implemented using highly-optimized matrix multiplication code.

■ Multi-Head Attention

Instead of performing a single attention function with large d_{model} dimensional queries, keys, and values, matrices are projected h times with different, learned linear projections to d_k , d_k , and d_v dimensions respectively (see Figure 2.10). This multi-head attention strategy allows the model to jointly attend to information from different representation subspaces at different positions, since attention operations can be performed in parallel. Output values are then concatenated and once again projected. Formally:

$$\text{MultiHead}(Q, K, V) := \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O,$$

$$\text{where } \text{head}_i := \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right).$$

In practice, due to the reduced dimension of each head, the total computational cost is similar to that of a single-head attention with full dimensionality.

■ Positional Encoding

Positional encoding is an important measure taken so that the transformer can make use of the order of the sequence without recurrence or convolution. It consists in the injection of some information about the relative or absolute position of tokens in the sequence:

$$\begin{aligned} \text{PE}_{(\text{pos}, 2i)} &:= \sin\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \\ \text{PE}_{(\text{pos}, 2i+1)} &:= \cos\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right), \end{aligned}$$

where pos is the position and i the dimension.

This function was chosen as it was hypothesized that the model would easily learn to attend by relative positions since, for any fixed offset k , $\text{PE}_{\text{pos}+k}$ can be represented as a linear function of PE_{pos} .

In practice, for each position in the sentence, positional embedding is computed as a d_{model} -dimensional vector which is added to the corresponding word embedding vector.

2.2.4 Pre-trained Language Models

Following the introduction of the transformer architecture, multiple language models based on transformers were proposed, with each one of them achieving new state-of-the-art performance. We can separate them into three different types: decoder-only, encoder-only, and encoder-decoder models, which respectively use either only the decoder part of the transformer architecture, only the encoder part, or both.

■ Decoder Models

In decoder models, for each word from the input sentence, the attention layers can only access the words positioned before it. These models are often called *auto-regressive* models. The main pretraining objective of decoder models is usually to predict the next word in the sentence. These models are best suited for tasks involving text generation.

GPT. [Radford et al. \[2018\]](#) propose the GPT model, whose name comes from their Generative Pre-Training procedure: using the transformer's masked multi-head self-attention, GPT is given a context sentence and asked to predict the following word. The model is then fine-tuned on specific tasks, sometimes using task-specific input transformations. They show that GPT outperforms the state of the art in 9 out of the 12 tasks studied.

GPT-2. [Radford et al. \[2019\]](#) suspect that the lack of generalization capabilities of language models is due to the prevalence of single task training on single domain datasets. They create WEBTEXT, a dataset that emphasizes document quality and only contains web pages that have been curated by humans. They then train four modified versions of GPT on a language modeling task on WEBTEXT. They show that their largest model, named GPT-2, achieves state-of-the-art results on 7 out of 8 language modeling datasets in a zero-shot setting, *i.e.*, without any fine-tuning.

■ Encoder Models

In encoder models, the attention layers can access every word of the input sentence. The pretraining objective typically consists in masking random words from the input and asking the model to predict the missing items. Encoder models are best suited for tasks requiring an understanding of the full sentence, such as sentence classification or named entity recognition.

BERT. [Devlin et al. \[2019\]](#) introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. BERT is intended to jointly condition on both left and right context in all layers in order to pretrain deep bidirectional representations. BERT is pre-trained using two unsupervised tasks: masked language modeling (MLM) and next sentence prediction (NSP). For MLM, since bidirectional conditioning would allow each word to indirectly “see itself”, some percentage of the input tokens is masked at random, and the model is then asked to predict those masked tokens. For NSP, when choosing the sentences A and B for each pretraining example, 50% of the time B is the actual next sentence that follows A (labeled as `IsNext`), and 50% of the time it is a random sentence from the corpus (labeled as `NotNext`). The model is then asked to predict whether sentence B is the sentence following sentence A. Finally, BERT is shown to obtain new state-of-the-art results on eleven natural language processing tasks.

RoBERTa. [Liu et al. \[2019\]](#) perform a replication study of BERT pretraining and evaluate the effects of many hyperparameters and training data size. They observe that BERT was noticeably undertrained and that correcting this issue enable it outperform more recent models. Their training modifications include increasing training duration, data size, or sequence length, removing the next sentence prediction objective, and dynamically varying the masking pattern. Their improved pretraining procedure, called RoBERTa, achieves new state-of-the-art results on multiple benchmarks, sometimes without multi-task finetuning or additional data.

■ Encoder-decoder Models

Encoder-decoder models attempt to leverage the strengths of both transformer components, through the incorporation of new pre-training objectives or architectural modifications. They are typically used for NLP tasks that involve understanding input sequences and generating output sequences, such as machine translation and text summarization.

XLNET. [Yang et al. \[2019\]](#) propose XLNET, a generalized autoregressive model that leverages the best of both decoder and encoder language models. First, unlike decoder models, XLNET maximizes the expected log-likelihood of a sequence w.r.t. all possible permutations of the factorization order, enabling it to capture bidirectional context. Secondly, unlike encoder models, XLNET does not rely on data corruption such as masking, and its autoregressive training objective removes the independence assumption that BERT makes. They show that XLNET consistently outperforms BERT on many tasks, *e.g.* language understanding, reading comprehension, and text classification.

BART. [Lewis et al. \[2020\]](#) present BART, a denoising encoder-decoder model based on Bidirectional and Auto-Regressive Transformers: it uses a standard transformer architecture, with a bidirectional encoder and an auto-regressive decoder. After corrupting text with an arbitrary noising function, the model learns to reconstruct the original text. Since arbitrary noising perturbations can be performed, the model is very flexible. BART achieves its best performance by both randomly shuffling the order of the original sentences and using a generalized version of token masking: it achieves state-of-the-art results on abstractive dialogue, question answering, summarization and machine translation.

T5. [Raffel et al. \[2020\]](#) explore transfer learning approaches and introduce a framework that translates all text-based NLP tasks in a text-to-text format. They filter the COMMONCRAWL dataset to produce the Colossal Clean Crawled Corpus (C4) dataset with a variation of BERT’s masked language modeling. They then pretrain a standard encoder-decoder transformer model on C4, which they name Text-to-Text Transfer Transformer (T5). They show that, after fine-tuning on downstream tasks, T5 achieves state-of-the-art performance on 18 out of 24 considered tasks.

2.2.5 Large Language Models

Overall, researchers often found that scaling pre-trained language models (e.g. scaling model or data size) led to an improved model performance on downstream tasks. [Kaplan et al. \[2020\]](#) study empirical scaling laws and observe consistent scalings of language model log-likelihood loss with non-embedding parameter count, dataset size, and optimized training computation. As pre-trained language models steadily increased in size, the research community coined the term “large language models” (LLM) to refer to them, starting from the 175B-parameter GPT-3 [[Brown et al., 2020](#)]. Among the large amount of proposed LLMs, most of which are decoder models, the most high-profile ones include CHATGPT¹, LAMDA [[Thoppilan et al., 2022](#)], INSTRUCTGPT [[Ouyang et al., 2022](#)], PALM [[Chowdhery et al., 2023](#)], GPT-4 [[Achiam et al., 2023](#)] and LLAMA2 [[Touvron et al., 2023b](#)].

Nowadays, LLMs are posing a significant impact on the AI community, and the advent of CHATGPT and GPT-4 leads to the rethinking of the possibilities of artificial general intelligence (AGI). [Altman \[2023\]](#) discusses the short-term and long-term plans to approach AGI, and [Bubeck et al. \[2023\]](#) have argued that GPT-4 might be considered as an early version of an AGI system. The research areas of AI in general are being revolutionized by the rapid progress of LLMs. In the field of NLP, LLMs can serve as general-purpose language task solvers (to some extent), and the research paradigm has been shifting towards the use of LLMs. However, it remains mysterious why emergent abilities occur in LLMs, instead of smaller language models. As a more general issue, there lacks a deep, detailed investigation of the key factors that contribute to the superior abilities of LLMs [[Wei et al., 2022a](#)].

We refer the reader to [Zhao et al. \[2023\]](#) for a more in-depth survey of LLMs.

¹<https://openai.com/blog/chatgpt/>

2.3 Evaluation and Meta-Evaluation

The question of evaluation has always been central to Natural Language Processing, as there is a strong need for procedures enabling researchers to track progress in a systematic and reliable manner. Different methods for evaluation have been proposed: notably, we can make a distinction between human and automatic evaluation methods. Moreover, the process of studying and comparing multiple evaluation strategies is called *meta-evaluation*. In this section, we provide some background for each evaluation type. Evaluation in the specific context of story generation is developed in [Chapter 3](#). We refer the reader to [\[Sai et al., 2022\]](#) for a survey of evaluation measures for evaluating Natural Language Generation in general.

2.3.1 Human Evaluation

Human evaluation has been tried in multiple setups for a variety of NLP tasks. When performing human evaluation, several factors have to be considered; we present them below. We tackle the human evaluation of story generation in [Section 3.2.1](#).

Evaluator Expertise. Depending on the demands of the task and the evaluation’s objective, the evaluators may be experts [\[Belz and Reiter, 2006\]](#), crowdsourced annotators [\[Callison-Burch, 2009\]](#), or even end users [\[See et al., 2019\]](#). If the review process speed is the main consideration, crowdsourced laborers are usually favored. In that case, however, it may be important to provide clear instructions, filter workers based on their prior performance, and conduct an extra layer of quality control, ideally with the assistance of one or two expert in-house annotators.

Evaluation Scale. The annotators are usually asked to score the output using the Likert scale, which is a fixed scale with each number denoting a distinct quality level [\[Likert, 1932\]](#), usually ranging from 1 to 5. Dynamic continuous scales, which enable the assessor to make more nuanced conclusions, have also been explored [\[Belz and Kow, 2011\]](#). Instead of providing a rating, the assessors can also be asked to give binary judgments, which may be favored over a rating scale to compel judges to render a definitive judgment as opposed to assigning an average rating (by selecting a score in the midpoint of the scale) [\[Horbach et al., 2020\]](#).

Providing references. In addition to the output generated by the system, one can supply the context (input) and a set of reference outputs. When the evaluation criteria can be reduced to a comparison of the information similarity between the two texts, references are useful: the adequacy of a generated translation may be assessed by comparing it to the reference. For other tasks, *e.g.* automatic summarization, providing the context might be the better option.

Relative or Absolute Evaluation. The generated candidate may be evaluated separately or w.r.t. to other candidates. In an individual evaluation setting, the candidate is provided an absolute rating for each desired criteria. However, in a comparative scenario, an annotator can be requested to rank the various candidates that are presented in a preferred order [\[Dušek](#)

et al., 2020] or asked to assess multiple candidates simultaneously [Novikova et al., 2017]. Alternatively, one might perform pairwise comparisons of two systems [Dhingra et al., 2019]: in such a setup, two systems are compared based on the number of times they were preferred (wins), not preferred (losses), and equally preferred (ties).

Providing an Explanation. The evaluators may also be asked to explain their choices, which they typically do by emphasizing the relevant passage that had an impact on the rating [Chaganty et al., 2018].

Inter-Annotator Agreement (IAA) : Typically, the same output is given to multiple evaluators; their evaluations are then aggregated to determine its final score. The aggregate can be calculated as a simple average or a weighted average, in which each annotator is given a weight determined by how well they have performed in the past or how well they agree with other annotators [Raykar et al., 2010]. It is generally desirable to obtain a high IAA, which can be estimated using Krippendorff’s alpha [Hayes and Krippendorff, 2007], Cohen’s Kappa [Cohen, 1960], or Fleiss’s Kappa [Fleiss, 1971]. On tasks involving more subjectivity (e.g. story evaluation or question answering), it is harder to achieve high IAAs [Amidei et al., 2018]. Moreover, a lower IAA may result from human error, inadequate setup or guidance, or ambiguous text [Sampson and Anna, 2008].

2.3.2 Automatic Evaluation

Since human evaluation can be expensive and time-consuming, multiple measures for automatic evaluation have been proposed for NLP tasks. Designing such measures requires choosing between different features, which we list below. We present our survey of automatic measures in Section 3.3.2.

Represented Criteria. Depending on the task, an automatic measure may need to capture a specific criterion. For instance, in machine translation, fluency and adequacy are important criteria [Graham et al., 2017], while text summarization places greater emphasis on informativeness and coherence [Hassel, 2004].

Input Type. Similarly to humans, automatic measures may require more input than just the candidate text. Providing context or reference outputs may enable more accurate scores. For machine translation, BLEU [Papineni et al., 2002] compares a candidate translation with a reference. For text summarization, BLANC [Vasilyev et al., 2020] does not need any reference summary but uses both the summary and the document to compute its score.

External Resources. When relevant, it is possible to make a measure use external resources, such as databases (e.g. WordNet [Miller, 1992], Yago [Suchanek et al., 2007], ConceptNet [Speer et al., 2017]), pre-processing tools (e.g. stemmers, lemmatizers, Part-of-Speech tagging), or pre-trained word embeddings (e.g. WORD2VEC [Mikolov et al., 2013b]). Some measures, such as NNEVAL [Sharif et al., 2018], use the scores provided by other measures as input.

Technique Used. There is a great number of possible designs for automatic measures, with varying degrees of complexity: some measures simply compute the length of the text [Fabbrri et al., 2021], while others evaluate barycentric distributions using the Wasserstein distance [Colombo et al., 2021].

2.3.3 Meta-Evaluation

Comparing Automatic and Human Evaluation. Several previous works have studied the relationship between automatic measures and human judgment. Zhang et al. [2004] find that the BLEU and NIST measures correlate poorly with human judgment for performing pairwise comparisons of Chinese machine translation systems and are difficult to interpret in terms of human criteria. Stent et al. [2005] study whether several automatic evaluation metrics are appropriate for evaluating fluency and adequacy for paraphrase generation and find that they are inadequate for fluency, and only barely adequate for adequacy. Callison-Burch et al. [2006] show that BLEU cannot be used for comparing systems which employ radically different strategies (especially comparing phrase-based statistical machine translation systems against systems that do not employ similar n -gram-based approaches). Ng and Abrecht [2015] show that the ROUGE measure is biased towards identifying lexical similarity when assessing the quality of a generated summary and improve on this by incorporating the use of word embeddings. Novikova et al. [2017] find that automatic metrics are particularly weak in distinguishing outputs of medium and good quality, which can be partially attributed to the fact that human judgements and automatic measure scores are given on different scales. They also show that measure performance is data- and system-specific. Reiter [2018] presents a structured review of the evidence on whether BLEU is a valid evaluation technique and concludes that the evidence does neither support using BLEU to evaluate other types of NLP systems (outside of machine translations), nor support using BLEU to evaluate individual texts rather than NLP systems. Ma et al. [2019] present the results of the WMT19 Metrics Shared Task in machine translation evaluation and find high system-level correlations but moderate to low segment-level correlations between automatic measures and human judgment. They also show that correlations are heavily affected by the underlying set of systems.

Task-Specific Meta-Evaluation. Meta-evaluation has been done in several subfields of NLP, e.g. image description [Elliott and Keller, 2014], dialogue response generation [Liu et al., 2016], question generation [Nema and Khapra, 2018], machine translation [Ma et al., 2019; Mathur et al., 2020], table-to-text generation [Dhingra et al., 2019], question answering [Chen et al., 2019], and summarization [Bhandari et al., 2020]. Crucially, at the start of our research, extensive meta-evaluation of Automatic Story Generation had yet to be done.

2.4 Conclusion

On the one hand, automatic story generation has been tried with a variety of rule-based techniques: structural models were the first story generation systems to be developed, followed by planning-based models. On the other hand, since the introduction of neural networks,

which are probabilistic in nature, neural-based language models have been demonstrating ever more impressive performance in multiple NLP tasks. As they are now being increasingly used for Automatic Story Generation, there is a growing need for the design of methods that allow for a reliable evaluation of ASG systems. In [Chapter 3](#), we set out to propose a new methodology for story evaluation.

Chapter 3

Methodology for Story Evaluation and ASG Meta-Evaluation

*“ ‘No, no! The adventures first,’ said
the Gryphon in an impatient tone:
‘explanations take such a dreadful
time.’ ”*

Alice’s Adventures in Wonderland,
Lewis Carroll

Contents

3.1 Introduction and Definitions	38
3.2 Designing New Criteria for Human Story Evaluation	39
3.2.1 Survey of Existing Methods of Human Story Evaluation in ASG	39
3.2.2 Approaches From the Social Sciences Literature	46
3.2.3 Our Proposed Human Criteria	48
3.3 Selecting Automatic Measures for Story Evaluation	49
3.3.1 Proposed Taxonomy of Automatic Evaluation Measures	49
3.3.2 Survey of Considered Automatic Measures	51
3.4 Proposed Framework for Meta-Evaluation	58
3.4.1 Related Work	58
3.4.2 Meta-Evaluation Strategies	58
3.4.3 Statistical Testing	61
3.5 Conclusion	62

Chapter 3 Abstract

In this chapter, we lay the groundwork for conducting a meta-evaluation of Automatic Story Generation. We begin by describing our chosen setting and providing definitions for the ASG and ASE tasks (Section 3.1). Then, we review the human evaluation methods used in the ASG literature (Section 3.2.1) and several approaches from the social sciences literature (Section 3.2.2). This allows us to propose an original, comprehensive set of six criteria (Section 3.2.3). We select and review a number of automatic evaluation measures designed for diverse NLP tasks (Section 3.3.2) and grouped them in different classes w.r.t. a newly-proposed taxonomy (Section 3.3.1). Finally, we describe our meta-evaluation framework, which we tailored for the ASG task (Section 3.4).

3.1 Introduction and Definitions

As explained in Section 2.3, reliable methods for the evaluation of NLP systems are crucial for comparing performances across models and identifying new research directions. Both human and automatic evaluation frameworks have been proposed for a variety of tasks, but not in the context of Automatic Story Generation (ASG). Indeed, while ASG systems have been exhibiting significant improvements in recent years, story evaluation has been lagging behind. This can be explained by the highly subjective nature of the task, coupled with the wide diversity of possible evaluation criteria.

Let us first describe our chosen setting. The ASG task commonly involves generating a story from a short sentence called a *prompt* [Alabdulkarim et al., 2021]. To avoid confusion, we may also call this prompt a *story-prompt*, so as to distinguish it from the Eval-Prompt, which will refer to the input prompt that is fed to a large language model (see Section 4.4.2). We define the tasks of Automatic Story Generation and Automatic Story Evaluation below.

Definition 5: Automatic Story Generation (ASG)

Given a story-prompt p_i , the ASG task consists in using a language model that takes p_i as its input for generating a story y_i .

Definition 6: Automatic Story Evaluation (ASE)

Given an automatic evaluation measure m (e.g. a scoring algorithm, a Large Language Model (LLM)...), a story-prompt p_i , and a story y_i , we define the ASE task as the production of an evaluation score $m(y_i)$.

In this chapter, we propose a new methodology for the evaluation of ASG, which will be used for our meta-evaluation experiments presented in Chapter 5. In Section 3.2, we describe our process for designing new criteria for human story evaluation, leading to the introduction of a comprehensive set of six human criteria: Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity. Then, in Section 3.3, we review a number of automatic

measures for text evaluation and propose a two-dimensional taxonomy that will allow for a more fine-grained analysis of automatic story evaluation. Finally, in [Section 3.4](#), we present a meta-evaluation framework that is tailored to the task of ASG, namely through the design of a specific meta-evaluation strategy called the overall correlation.

3.2 Designing New Criteria for Human Story Evaluation

In Natural Language Processing, human evaluation has always been the gold standard for any evaluation procedure, and even more so for ASG where it is the only reliable evaluation method. We started by surveying how it was performed in the ASG literature, and found out that there was a glaring lack of a unified procedure for human story evaluation.

3.2.1 Survey of Existing Methods of Human Story Evaluation in ASG

[Lee et al. \[2019\]](#) observe that there is little agreement as to how natural language generation systems should be evaluated, with a particularly high degree of variation in the way that human evaluation is carried out. They provide an overview of how human evaluation is currently conducted and present a set of best practices, grounded in the literature. In particular, they advise to define separate and precise criteria for human evaluation to make it as accurate as possible. We argue that this is especially relevant to ASG, as evaluating a story has its fair share of subjectivity. However, we find that there is also little consensus in evaluation of ASG performance. We list below the diverse story evaluation criteria we found in the literature, ordered in chronological order of occurrence. [Table 3.1](#) and [Table 3.2](#) provide a synoptic overview of those criteria.

Coherence:

- [Khalifa et al. \[2017\]](#) perform a user study where they compare two pieces of text and ask “Which text is more coherent?”;
- For [Xu et al. \[2018\]](#), coherence evaluates “whether the generated story is coherent”;
- [Yao et al. \[2019\]](#) conduct pairwise comparisons on coherence, defined as “whether the story is logically consistent and coherent”;
- [Jhamtani and Berg-Kirkpatrick \[2020\]](#) evaluate story samples w.r.t. “coherence, which measures the logical and coherent narrative flow in a story”;
- [Akoury et al. \[2020\]](#) ask STORIUM users to evaluate generated text on a 5-point Likert scale w.r.t. “coherence (logical ordering of sentences)”;
- [Wang et al. \[2020\]](#) define coherence as “whether a story is logically consistent or describes a possible scenario”;
- [Guan et al. \[2021a\]](#) define coherence as “inter-sentence relatedness, causal and temporal dependencies”;
- [Wilmot and Keller \[2021\]](#) ask Mechanical Turk workers to rank five candidate stories from best to worst w.r.t. “coherence (internal consistency of characters, places, actions and plot development)”;

Interestingness:

- [Khalifa et al. \[2017\]](#) perform a user study where they compare two pieces of text and ask “Which text is more interesting?”;
- [Yao et al. \[2019\]](#) conduct pairwise comparisons on interestingness, defined as “whether the story is interesting”;

Grammar:

- [Khalifa et al. \[2017\]](#) perform a user study where they compare two pieces of text and ask “Which text is more grammatically correct?”
- [Guan et al. \[2019\]](#) define grammar as “whether an ending is natural and fluent”;
- [Guan et al. \[2020\]](#) define grammaticality as “whether a story is natural and fluent”;

Sentence Preference:

- [Clark et al. \[2018\]](#) present human judges with a short excerpt from a story and two generated sentences, one generated by their ENGEN model and one generated by their entity-unaware S2SA model. They then ask them to “choose a sentence to continue the story” and to briefly explain why they made the choice they did;
- [Clark and Smith \[2021\]](#) ask human writers to choose between suggestions of story completion that are generated by different models;

Faithfulness:

- [Peng et al. \[2018\]](#) define faithfulness as “whether the generated stories follow a given storyline of five words”;
- [Wang et al. \[2020\]](#) define faithfulness as “whether a story is faithful to a given story beginning and ending”;

Relevance / Fidelity:

- [Fan et al. \[2018\]](#) conduct a triple pairing task: they use each model to generate stories based on held-out prompts from the test set. Then, groups of three stories are presented to the human judges. The stories and their corresponding prompts are shuffled, and human evaluators are asked to select the correct pairing for all three prompts. This evaluates to which extent the model is able to generate a story that is relevant to the prompt;
- [Yao et al. \[2019\]](#) conduct pairwise comparisons on fidelity, defined as “whether the story is on-topic with the given title”;
- [Jhamtani and Berg-Kirkpatrick \[2020\]](#) evaluate story samples w.r.t. “fidelity to title, which measures the degree to which the story is relevant to the given title”;
- [Akoury et al. \[2020\]](#) ask STORIUM users to evaluate generated text on a 5-point Likert scale w.r.t. “relevance (fit with the current story)”;
- [Goldfarb-Tarrant et al. \[2020\]](#) define relevance as “whether stories both relate to the given prompt and remain on topic for the duration of the story”;
- [Bai et al. \[2021\]](#) ask four native speakers to compare the generations of different systems in pairs w.r.t. topic relevance, without further definition;

Fluency:

- [Xu et al. \[2018\]](#) define fluency as “whether each sentence in the generated story is correct in grammar”;
- [Akoury et al. \[2020\]](#) ask STORIUM users to evaluate generated text on a 5-point Likert scale w.r.t. “fluency (judgment of grammaticality)”;
- [Guan et al. \[2021a\]](#) define fluency as “a measure of intra-sentence linguistic quality and grammatical correctness”;
- [Bai et al. \[2021\]](#) ask four native speakers to compare the generations of different systems in pairs w.r.t. fluency, without further definition;
- [Wilmot and Keller \[2021\]](#) ask Mechanical Turk workers to rank five candidate stories from best to worst w.r.t. “fluency (how well written the story is, including its use of expressive language)”;
- [Pascual et al. \[2021\]](#) ask human judges to rate stories on a 1–7 Likert scale w.r.t. “fluency (the text can be understood easily, well, and quickly)”;

Logicality:

- [Guan et al. \[2019\]](#) define logicality as “whether an ending is reasonable and coherent with the story context in logic”;
- [Guan et al. \[2020\]](#) define logicality as “whether a story is coherent to the given beginning and reasonable in terms of causal and temporal dependencies in the context”;
- [Pascual et al. \[2021\]](#) ask human judges to rate stories on a 1–7 Likert scale w.r.t. “logical consistency (there are no contradictions in the text)”;

Overall Preference:

- [Yao et al. \[2019\]](#) conduct pairwise comparisons on overall user preference, defined as “how do users like the story”;
- [Fan et al. \[2019\]](#) show two different stories (that were generated based on the same human-written prompt) to human judges, who do not see the prompt. The evaluators are then asked to mark which story they prefer;
- [Brahman and Chaturvedi \[2020\]](#) ask workers to evaluate pairs of stories on a 0–3 scale (3 being very good) on content quality, *i.e.*, “to indicate whether a story is fluent, logically coherent, and on-topic (related to the given title)”;
- [Wang et al. \[2020\]](#) define overall preference as “whether the human judge like a sequence of sentences as a story”;
- [Bai et al. \[2021\]](#) ask four native speakers to compare the generations of different systems in pairs w.r.t. overall preference;

Likability:

- [Akoury et al. \[2020\]](#) ask STORIUM users to evaluate generated text on a 5-point Likert scale w.r.t. “likability (subjective assessment of enjoyability)”;

Overall Quality:

- Goldfarb-Tarrant et al. [2020] define overall quality to be “a combination of coherence, interestingness, and relevance, similarly to most other story generation works”;
- Wilmot and Keller [2021] ask Mechanical Turk workers to rank five candidate stories from best to worst w.r.t. “overall (a subjective judgement on their overall impression of the story)”;
- Guan et al. [2021b] decide the overall quality of a story by “summarizing multiple separate criteria: a story should get punishment in points if it contains errors such as repetitive plots, unrelated events and conflicting logic, or globally chaotic scenes”;

Outline Utilisation:

- Rashkin et al. [2020] give raters two paragraphs, each generated by different models, and ask them to select which is utilizing the provided outline better;

Narrative Flow:

- Rashkin et al. [2020] give raters a generated paragraph (with the previous paragraph as context). They are asked to rate on a scale from 1 to 5 how much the paragraph: (a) repeats content from the previous paragraph, (b) transitions naturally from the previous paragraph, and (c) stays relevant and on-topic throughout the paragraph;

Ordering:

- Rashkin et al. [2020] give raters a pair of consecutive generated paragraphs, presented in a random order, and ask them to attempt to decipher the order. The intuition is that, if the model output is very well-structured, then it should be easier for humans to decipher the order;

Emotion Faithfulness:

- Brahman and Chaturvedi [2020] ask workers to evaluate pairs of stories on a 0–3 scale (3 being very good) on emotion faithfulness, *i.e.*, “to assess whether it follows the desired emotion arc for the protagonist”;

Content Quality:

- Brahman and Chaturvedi [2020] ask workers to evaluate pairs of stories on a 0–3 scale (3 being very good) on content quality, *i.e.*, “to indicate whether a story is fluent, logically coherent, and on-topic (related to the given title)”;

Human Plausibility:

- Ghazarian et al. [2021] ask human annotators to rate stories on a scale of 0 to 5 (from “not at all plausible” to “completely plausible”);

Ending Quality:

- Bai et al. [2021] ask four native speakers to compare the generations of different systems in pairs w.r.t. ending quality, without further definition;

Naturalness:

- Pascual et al. [2021] ask human judges to rate stories on a 1–7 Likert scale w.r.t. “naturalness (the text seems to have been written by a human)”;

Takeaways. Many of these criteria are not specific to Automatic Story Generation (grammar, fluency, overall preference, overall quality, content quality, human plausibility, naturalness), overlap with one another (faithfulness and relevance; coherence, logicity and narrative flow) or are ascribed to a specific setting (sentence preference, outline utilisation, ordering, emotion faithfulness, ending quality). Furthermore, evaluation protocols mostly use only two or three criteria, which is not enough to grasp all aspects of a task as complex as ASG. They also do not associate Likert scales with explicit descriptions, even though such descriptions could reduce the subjectivity of the labelling process.

In order to propose a more comprehensive set of criteria for human evaluation of story generation, we delved into the social sciences literature.

Criterion	Source	Question / Definition	Evaluation Type
Coherence	Khalifa et al. [2017]	“Which text is more coherent?”	Pairwise Comparison
	Xu et al. [2018]	“Whether the generated story is coherent”	10-point Likert Scale
	Yao et al. [2019]	“Whether the story is logically consistent and coherent”	Pairwise Comparison
	Jhamtani and Berg-Kirkpatrick [2020]	“Logical and coherent narrative flow”	Pairwise Comparison
	Akoury et al. [2020]	“Logical ordering of sentences”	5-point Likert Scale
	Wang et al. [2020]	“Whether a story is logically consistent or describes a possible scenario”	Pairwise Comparison
	Guan et al. [2021a]	“Inter-sentence relatedness, causal and temporal dependencies”	Pairwise Comparison
	Wilmot and Keller [2021]	“Internal consistency of characters, places, actions, and plot development”	5-fold Ranking
Interestingness	Khalifa et al. [2017]	“Which text is more interesting?”	Pairwise Comparison
	Yao et al. [2019]	“Whether the story is interesting”	Pairwise Comparison
Grammar	Khalifa et al. [2017]	“Which text is more grammatically correct?”	Pairwise Comparison
	Guan et al. [2019]	“Whether an ending is natural and fluent”	3-point Likert Scale
	Guan et al. [2020]	“Whether an story is natural and fluent”	Pairwise Comparison
Sentence Preference	Clark et al. [2018]	“Choose a sentence to continue the story”	Pairwise Comparison
	Clark and Smith [2021]	Choose between suggestions of story completion	Pairwise Comparison
Faithfulness	Peng et al. [2018]	“Whether the generated stories follow a given storyline of five words”	5-point Likert Scale
	Wang et al. [2020]	“Whether a story is faithful to a given story beginning and ending”	Pairwise Comparison
Relevance / Fidelity	Fan et al. [2018]	Select the correct prompt-story pairing	Triple Pairing Task
	Yao et al. [2019]	“Whether the story is on-topic with the given title”	Pairwise Comparison
	Jhamtani and Berg-Kirkpatrick [2020]	“The degree to which the story is relevant to the given title”	Pairwise Comparison
	Akoury et al. [2020]	“Fit with the current story”	5-point Likert Scale
	Goldfarb-Tarrant et al. [2020]	“Whether stories both relate to the given prompt [...]”	5-point Likert Scale
	Bai et al. [2021]	N/A	Pairwise Comparison
Fluency	Xu et al. [2018]	“Whether each sentence is correct in grammar”	10-point Likert Scale
	Akoury et al. [2020]	“Judgment of grammaticality”	5-point Likert Scale
	Guan et al. [2021a]	“Intra-sentence linguistic quality and grammatical correctness”	Pairwise Comparison
	Bai et al. [2021]	N/A	Pairwise Comparison
	Wilmot and Keller [2021]	“How well written the story is”	5-fold Ranking
	Pascual et al. [2021]	“The text can be understood easily, well, and quickly”	7-point Likert Scale

Table 3.1: Overview of human story evaluation practices in ASG (Part 1 of 2).

Criterion	Source	Question / Definition	Evaluation Type
Logicity	Guan et al. [2019] Guan et al. [2020] Pascual et al. [2021]	“Whether an ending is reasonable and coherent with the story context in logic” “Whether a story is coherent to the given beginning [...]” “No contradictions in the text”	3-point Likert Scale Pairwise Comparison 7-point Likert Scale
Overall Preference	Yao et al. [2019] Fan et al. [2019] Brahman and Chaturvedi [2020] Wang et al. [2020] Bai et al. [2021]	“How do users like the story” Mark the preferred story “Whether a story is fluent, logically coherent, and on-topic [...]” “Whether the human judge like a sequence of sentences as a story” N/A	Pairwise Comparison Pairwise Comparison 4-point Likert Scale Pairwise Comparison Pairwise Comparison
Likability	Akoury et al. [2020]	“Subjective assessment of enjoyability”	5-point Likert Scale
Overall Quality	Goldfarb-Tarrant et al. [2020] Wilmot and Keller [2021] Guan et al. [2021b]	“A combination of coherence, interestingness, and relevance” “A subjective judgement on their overall impression of the story” “A story should get punishment in points if it contains errors [...]”	5-point Likert Scale 5-fold Ranking 5-point Likert Scale
Outline Utilisation	Rashkin et al. [2020]	Select the paragraph that uses the provided outline better	Pairwise Comparison
Narrative Flow	Rashkin et al. [2020]	3 ratings: repetitions, natural transitions and relevance	5-point Likert Scale
Ordering	Rashkin et al. [2020]	Retrieve the original ordering of shuffled paragraphs	Accuracy
Emotion Faithfulness	Brahman et al. [2020]	“Whether it follows the desired emotion arc for the protagonist”	4-point Likert Scale
Content Quality	Brahman et al. [2020]	Whether a story is fluent, logically coherent, and on-topic	4-point Likert Scale
Human Plausibility	Ghazarian et al. [2021]	From “Not at all plausible” to “Completely plausible”	6-point Likert Scale
Ending Quality	Bai et al. [2021]	N/A	Pairwise Comparison
Naturalness	Pascual et al. [2021]	“The text seems to have been written by a human”	7-point Likert Scale

Table 3.2: Overview of human story evaluation practices in ASG (Part 2 of 2).

3.2.2 Approaches From the Social Sciences Literature

The question “What makes a good story?” has been tackled from many different angles, and it would be impossible to review all the literature on this topic. In this section, we present a few notable works which propose a framework for describing the quality of a story.

High Point Structure, Episodic Grammars, and Dependency Analysis. McCabe and Peterson [1984] select three different methods for examining the structure of a story, each of them emphasizing a distinct aspect: (1) the high point structure [Labov and Waletzky, 1997] focuses on affective information and considers that emotional highs or crisis events are the key components of a story; (2) episodic grammars (also called story grammars) [Rumelhart, 1975] liken stories to a succession of problem-solving episodes where the protagonist attempts to accomplish a specific goal; and (3) the dependency analysis [Deese, 1978] associates story quality with linguistic complexity, and especially looks at whether propositions are coordinate or subordinate to one another.

Stories were then scored on 6- to 8-point Likert scales, with precise descriptions of each rating, according to how well they realized good structure in each system. They show that the three analyses cover mostly independent aspects of storytelling, despite a small overlap between the episodic and high point methods. Most notably, they find that the best narratives are highly-rated in at least two of the three dimensions, while the worst stories are poorly-rated in all aspects. In other words, people generally enjoy stories that involve (1) engaging emotional commentary, (2) complex problem-solving, and (3) detailed descriptions of characters and events.

Empedocles’s Four Elements. Dickman [2003] finds that the best definition for a story is: “a fact wrapped in an emotion that can compel us to take action and so transform the world around us.” They note that the four fundamental elements of any effective story are passion, hero, antagonist, and transformation. They ground their analysis in the framework of four elements (fire, earth, air, and water) developed by the philosopher and poet Empedocles, student of Pythagoras, putting forth that previous work showed how those four elements could also reflect inner psychological states [Hadot, 2002].

They refer to the emotion that encircles the main idea and captivates the audience’s attention as passion, which is associated with Empedocles’s fire. A story is rarely a cohesive whole when it is first presented to an audience: it is made up of independent characters with various needs and goals. Passion is the fire that grabs the readers’ attention and pulls them into the narrative.

The earth element is linked to the hero: it symbolizes the manner in which the narrative is anchored in reality. Indeed, the point of view of the hero is our point of entry to the story: it must be sufficiently grounded to support the plot while remaining relatable enough for the readers to follow. Throughout the story, the hero serves as both a guide and a stand-in.

The antagonist is associated with air. They give the story energy and vitality, since the audience considers a story to be flat if there are no barriers. Engaging with the antagonist fosters an atmosphere that lends credibility and intrigue to the story. The antagonist serves

as the personification of a greater problem: they crystallize many of the abstract aspects of the plot. In most excellent stories, a strong antagonist provides direction and purpose to the hero's actions, as well as to the audience's attention.

Last, when a tale is conveyed correctly, transformation happens naturally as a consequence. When the hero sets out to conquer the challenges they face, they become a vector of transformation for the world. Since water can change the most out of the four elements, it is the one that is associated with transformation. As the hero emerges from the flames of hell as a better person, the audience is satisfied and learns to moving forward from the negative toward the positive.

Cognitive and Emotional Interest. Bae et al. [2021] present an incipient framework for measuring story interestingness that revolves around two types of factors: cognitive interest and emotional interest. The cognitive factors include four components: goal, novelty, inference, and schema violation, while the emotional factors include four other elements: empathy, external emotions, humor, and outcome valence.

- Cognitive Factors:

1. The goal is the protagonist's desire, which drive a story forward. The protagonist fights to fulfill their ambitions, which often run contrary to the antagonist's own goals. This generally leads to conflicts that are central to the story. The protagonist's objective must be significant and challenging to accomplish in order for the story to be engaging. A strong logline that has the power to grab the audience's interest usually sums up the goals or objectives of the protagonist in one sentence.
2. Novelty is a key component of narrative intelligence, as it helps for many tasks such as characterization, narration, and thematization. Good stories very often exhibit high novelty, as it tends to raise the interest of the reader.
3. Predictive inference describes the process of predicting what will happen next in a story, or how a specific event happened. It greatly influences the interestingness of a story and requires a careful handling of the reader's prior information and level of uncertainty. Generally, readers prefer stories with high postdictability, *i.e.*, stories that are crafted in such a way that we can at some point put all the parts together as a coherent whole, including inferences and the story's conclusion. This feature is often a result of skillful foreshadowing, which is a device.
4. Readers often rely on schema to understand a story, where a schema refers to "a data structure for representing the generic concepts stored in memory". Schema violation describes the process of deviating from common schema on purpose, leading to unexpected events that may spark greater interest to the story. However, schema violation requires proper resolution to be effective.

- Emotional Factors:

1. "Cognitive" (mental perspective-taking) or "emotional" (raw sharing of emotions) empathy can both be developed through storytelling. Unlike sympathy, empathy in stories is linked to positive feelings, *e.g.* happiness, contentment, exaltation. Those

are essential to the reader’s enjoyment of a story. Empathy for fictional characters includes identification with their intentions and aspirations on an internal level as well as recollections of their past experience.

2. Common external emotions of literary response include curiosity, suspense, and surprise, which are often associated to story schema. Crafting a good narrative structure often enables the author to elicit desired external emotions in the reader; they can be easily connected to the cognitive factors of interest.
3. Humor is a positive experience that frequently coexists with other positive feelings such as playfulness and delight. Humor is known to be very subjective: a humor performer’s intentions can be interpreted in wildly different ways. Three dimensions are generally addressed by theories on humor: superiority, incongruity, and relief. Just as surprise benefits greatly from the appropriate resolution of incongruity, humor also greatly benefits from incongruity and its resolution. Humor appears to be a sufficient rather than a necessary component of story interest, which makes its inclusion in the list of emotional factors more debatable.
4. The extent to which readers enjoy a story is often related to its outcome valence: good or bad, happy or sad. Outcome valence is generally a consequence of the proper resolution of the story’s unexpectedness, which are mentioned in the cognitive factors of story interestingness.

In the next section, we explain our decision process for coming up with a fixed set of criteria for human evaluation of ASG.

3.2.3 Our Proposed Human Criteria

Deciding on a fixed set of criteria for as subjective a task as story evaluation is undoubtedly a delicate endeavour. Upon reviewing the ASG and social sciences literature, we focused on several components that seemed to be frequently identified as reliable indicators of story quality. From the ASG literature, we reckoned that a vast majority of the observed criteria could be boiled down to two: Relevance, which is specific to our chosen setting of generating a story from a prompt, and Coherence, which is almost omnipresent when accounting for its different names. For the rest of our criteria, we mainly relied on the cognitive and emotional factors from [Bae et al. \[2021\]](#); for each of them, we looked at how they related to the frameworks proposed by [McCabe and Peterson \[1984\]](#) and [Dickman \[2003\]](#), as well as other sources.

After a careful review, we condensed the frameworks presented in [Section 3.2.2](#) into a comprehensive set of criteria for story evaluation:

1. **Relevance** (RE): “how well the story matches its prompt”, used in [Akoury et al. \[2020\]](#); [Bai et al. \[2021\]](#); [Fan et al. \[2018\]](#); [Goldfarb-Tarrant et al. \[2020\]](#); [Jhamtani and Berg-Kirkpatrick \[2020\]](#); [Yao et al. \[2019\]](#);
2. **Coherence** (CH): “how much the story makes sense”, used in [Akoury et al. \[2020\]](#); [Guan et al. \[2021a\]](#); [Jhamtani and Berg-Kirkpatrick \[2020\]](#); [Khalifa et al. \[2017\]](#); [Wang et al. \[2020\]](#); [Wilmot and Keller \[2021\]](#); [Xu et al. \[2018\]](#); [Yao et al. \[2019\]](#);

3. **Empathy** (EM): “how well the reader understood the character’s emotions”, derived from the importance of emotional commentary [McCabe and Peterson, 1984], passion [Dickman, 2003], and empathy [Bae et al., 2021; Keen, 2007];
4. **Surprise** (SU): “how surprising the end of the story was”, derived from the importance of schema violation, or unexpectedness [Bae et al., 2021; Schank, 1979], postdictability [Behrooz et al., 2019], and novelty [Randall, 1999];
5. **Engagement** (EG): “how much the reader engaged with the story”; a more subjective criterion associated with projecting volitive modality (making the reader formulate a subjective judgment and express a desire to see something accomplished) [Toolan, 2012] and story outcome, which is an underlying cause of story liking [Iran-Nejad, 1987];
6. **Complexity** (CX): “how elaborate the story is”; derived from the importance of detailed descriptions and sophisticated problem-solving [McCabe and Peterson, 1984] and good world-building [Roine, 2016].

The four last criteria are an original contribution and were designed to evaluate story features that are different and, ideally, independent from the first two criteria (RE and CH), which were already commonly used in the ASG literature.

In the next section, we focus on the automatic methods of ASG evaluation.

3.3 Selecting Automatic Measures for Story Evaluation

While human evaluation is unmatched when it comes to evaluation quality, it requires skilled annotators and elaborate guidelines, which makes it a time-consuming and expensive task. That is why automatic evaluation measures (such as BLEU [Papineni et al., 2002]) were developed.

The purpose of an evaluation measure is to assign a score to the outputs generated by a system and provide a quick and easy means of comparing different systems and tracking progress. Many evaluation measures compare a candidate text with a reference, which make them poorly adapted in the context of Automatic Story Generation: a given prompt can give birth to a myriad of different stories that remain relevant to it. And yet, most of the research work in ASG relies on such such as BLEU or ROUGE [Lin, 2004], despite the fact that many other automatic evaluation measures have been developed.

Since no Automatic Story Evaluation measure has been proposed, we find important to study whether there exists better candidates for ASE than the omnipresent BLEU and ROUGE. We selected a number of automatic measures for that purpose, which we present below.

3.3.1 Proposed Taxonomy of Automatic Evaluation Measures

We propose a taxonomy which classifies automatic measures along two independent axes, which we later use in our meta-evaluation benchmark in order to identify potential trends across the different categories of evaluation measures. First, an evaluation measure can belong to either of the two following categories:

- *Reference-based* (Ξ): it evaluates a candidate text by comparing it to a reference text. In our ASG setting, that would mean comparing a candidate story with a reference story written for the same prompt;
- *Reference-free* ($\sqsupset$): it relies only on the candidate text. In our ASG setting, this would mean evaluating only the candidate story (and, possibly, the prompt).

Second, irrespective of their being reference-based or reference-free, an evaluation measure can also belong to one of the three following categories:

- *String-based* ($\S$): it evaluates the textual representation of the inputs. Such simple measures are typically unable to handle synonyms or paraphrases;
- *Embedding-based* (ϵ): it relies on word embeddings, *e.g.* WORD2VEC [Mikolov et al., 2013a,b], or contextualized embeddings, *e.g.* obtained from BERT [Devlin et al., 2019];
- *Model-based* (Δ): it leverages regression or pre-trained language models to return a score.

In particular, for the task of Automatic Story Evaluation, one would expect reference-free measures to be more adapted since stories can be very different (both from a syntactic and semantic point of view) and yet of similar “quality”, even if they were generated from the same prompt.

	Reference-based (Ξ)	Reference-free ($\sqsupset$)
String-based ($\S$)	BLEU [Papineni et al., 2002]	COVERAGE [Grusky et al., 2018]
	ROUGE [Lin, 2004]	DENSITY [Grusky et al., 2018]
	METEOR [Banerjee and Lavie, 2005]	COMPRESSION [Grusky et al., 2018]
	CHRF [Popović, 2015]	TEXT LENGTH [Fabbri et al., 2021]
	CIDeR [Vedantam et al., 2015]	NOVELTY [Fabbri et al., 2021]
		REPETITION [Fabbri et al., 2021]
Embedding-based (ϵ)	ROUGE-WE [Ng and Abrecht, 2015]	
	BERTSCORE [Zhang et al., 2020]	
	MOVERSCORE [Zhao et al., 2019]	SUPERT [Gao et al., 2020]
	BARYSCORE [Colombo et al., 2021]	
	DEPTHSCORE [Staerman et al., 2024]	
Model-based (Δ)	S3 [Peyrard et al., 2017]	
	SUMMAQA [Scialom et al., 2019]	BLANC [Vasilyev et al., 2020]
	INFOLM [Colombo et al., 2022a]	
	BARTSCORE [Yuan et al., 2021]	

Table 3.3: Taxonomy of the automatic measures considered in our study with symbols for easier identification. Note: BARTSCORE was designed to be either reference-based or reference-free depending on the setting.

A synoptic view of the automatic measures considered in our study can be found in Table 3.3. We made sure to select measures that belong to each category of our taxonomy.

3.3.2 Survey of Considered Automatic Measures

For a more comprehensive account of the automatic evaluation measures used in NLP, we refer the reader to [Chauhan and Daniel \[2023\]](#); [Sai et al. \[2022\]](#).

BLEU. [Papineni et al. \[2002\]](#) introduce BLEU (BiLingual Evaluation Understudy), by far the most widely used automatic evaluation measure in NLP. It is a precision-based measure that computes the n -gram overlap between the reference and the hypothesis. In particular, BLEU is the ratio of the number of overlapping n -grams to the total number of n -grams in the hypothesis. This precision is computed separately for different values of n as shown below:

$$\text{Precision}_n = \frac{\sum_{p \in \text{hypotheses}} \sum_{n\text{-gram} \in p} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{p \in \text{hypotheses}} \sum_{n\text{-gram} \in p} \text{Count}(n\text{-gram})},$$

where $\text{Count}_{\text{clip}}(n\text{-gram})$ is clipped by the maximum number of times the given n -gram appears in any of the corresponding reference sentences. Once the above precision is computed for different values of n , a final BLEU-N score is computed as a weighted combination (in this case, a geometric mean) of all the Precision_n scores for $n = 1, \dots, N$. Since precision depends only on the length of the hypothesis and not on the length of the reference, an NLG system can exploit the measure and acquire high scores by producing only a few matching words or n -grams as the hypothesis. To discourage such short meaningless hypotheses, a brevity penalty term, BP, is added to the formula. The final formula commonly used today is:

$$\text{BLEU-N} = \text{BP} \cdot \exp \left(\sum_{n=1}^N u_n \log(\text{Precision}_n) \right), \quad \text{where} \quad \text{BP} = \begin{cases} 1 & \text{if } |p| > |r| \\ \exp \left(1 - \frac{|r|}{|p|} \right) & \text{otherwise} \end{cases},$$

where the u_n are the weights of the different n -gram precisions, such that $\sum_{n=1}^N u_n = 1$ (usually, $u_n = \frac{1}{N}$).

ROUGE. [Lin \[2004\]](#) proposes ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which includes a set of variants: ROUGE- n , ROUGE-L, ROUGE-W, and ROUGE-S. ROUGE- n is similar to the Precision_n term of BLEU-N in counting the n -gram matches between the hypothesis and reference, so we will henceforth refer to it as simply ROUGE. The main difference with BLEU is that it is a recall-based measure whereas BLEU is precision-based:

$$\text{ROUGE-}n = \frac{\sum_{s_r \in \text{references}} \sum_{n\text{-gram} \in s_r} \text{Count}_{\text{match}}(n\text{-gram})}{\sum_{s_r \in \text{references}} \sum_{n\text{-gram} \in s_r} \text{Count}(n\text{-gram})}.$$

METEOR. [Banerjee and Lavie \[2005\]](#) point out that there are two major drawbacks of BLEU: (1) it does not take recall into account, and (2) it only allows exact n -gram matching. They then introduce METEOR (Metric for Evaluation of Translation with Explicit Ordering), which is based on a F-score and uses a relaxed matching criteria. In particular, even if a unigram in the hypothesis does not have an exact surface level match with a unigram in the reference but is still equivalent to it (e.g. a synonym), then METEOR considers this as a matched unigram.

More specifically, it first performs exact word (unigram) mapping, followed by stemmed word matching, and finally synonym and paraphrase matching. It then computes the F-score using this relaxed matching strategy (P and R stand for *Precision* and *Recall* respectively):

$$\text{F-score} = \frac{10PR}{R + 9P},$$

$$\text{where } P = \frac{\#\{\text{mapped unigrams}\}}{\#\{\text{unigrams in candidate}\}} \quad \text{and} \quad R = \frac{\#\{\text{mapped unigrams}\}}{\#\{\text{unigrams in reference}\}}.$$

Since METEOR only considers unigram matches (as opposed to n -gram matches), it seeks to reward longer contiguous matches using a penalty term known as “fragmentation penalty”. To compute this, “chunks” of matches are identified in the hypothesis, where contiguous hypothesis unigrams that are mapped to contiguous unigrams in a reference can be grouped together into one chunk. Therefore longer n -gram matches lead to fewer number of chunks. The fewest possible number of chunks a hypothesis can have is used to compute the fragmentation penalty used by METEOR as:

$$\text{METEOR} = \text{F-score} \cdot (1 - \text{Penalty}), \quad \text{where} \quad \text{Penalty} = 0.5 \cdot \left(\frac{\#\{\text{chunks}\}}{\#\{\text{matched unigrams}\}} \right)^3.$$

CHRF. Popović [2015] propose the use of character n -gram F-score (CHRF) for the automatic evaluation of machine translation output. They argue that it is a simple measure that takes into account morpho-syntactic phenomena, does not require any additional tools and/or knowledge sources, is absolutely language independent, and is also tokenization independent. They define their measure as follows:

$$\text{CHRF} \beta := (1 + \beta^2) \frac{\text{CHRP} \cdot \text{CHRR}}{\beta^2 \cdot \text{CHRP} + \text{CHRR}},$$

where CHRP and CHRR stand for character n -gram precision and recall arithmetically averaged over all n -grams, and β is a parameter which assigns β times more importance to recall than precision — if $\beta = 1$, they have the same weight.

They compute correlations with human rankings on WMT12, WMT13, and WMT14 data, and find that CHRF outperforms other measures for a majority of texts at both system-level and segment-level.

CIDER. Vedantam et al. [2015] introduce CIDER, an automatic measure that weighs each n -gram in a sentence based on its frequency in the corpus and in the reference set of the particular instance, using term-frequency and inverse-document-frequency (TF-IDF). It was first proposed in the context of Image Captioning where each image is accompanied by multiple reference captions. It is based on the premise that n -grams that are relevant to an image would occur frequently in its set of reference captions. However, n -grams that appear frequently in the entire dataset (*i.e.*, in the reference captions of different images) are less likely to be relevant and hence they are assigned a lower weight using inverse-document-frequency

(IDF) term. To be more precise, the TF-IDF weight, $g_{n_k}(s)$, for each n -gram k in caption s_i are computed as follows:

$$g_{n_k}(s) = \frac{t_k(s)}{\sum_{l \in V_n} t_l(s)} \log \left(\frac{|I|}{\sum_{i \in I} \min(1, \sum_{r \in R_i} t_k(r))} \right),$$

where V_n is the vocabulary of all n -grams, g_{n_k} refers to the weight assigned to n -gram k , $t_k(s)$ is the number of times k appears in s , I is the set of all images, and R_i corresponds to the set of references for image i .

CIDER first stems the words in hypothesis and references and represents each sentence as a set of n -grams. It then calculates weights for each n -gram using TF-IDF. Using these TF-IDF weights of all the n -grams of length n , vectors $g_n(s)$ are formed for each caption s . CIDER_n is calculated as the average cosine similarity between hypothesis and references. The final CIDER score is the weighted average of CIDER_n for $n = 1, 2, 3, 4$:

$$\text{CIDER}(p, R) = \sum_{n=1}^N W_n \text{CIDER}_n(p, R), \quad \text{where} \quad \text{CIDER}_n(p, R) = \frac{1}{|R|} \sum_{r \in R} \frac{g_n(p) \cdot g_n(r)}{\|g_n(p)\| \|g_n(r)\|}$$

where $g_n(p)$ is a vector of the TF-IDF weights in p corresponding to all n -grams of length n , and uniform weights $W_n = \frac{1}{n}$ are usually used with $N = 4$.

ROUGE-WE. Ng and Abrecht [2015] integrate word embeddings into ROUGE in order to tackle its bias towards lexical similarities. They define a new similarity function

$$f_{\text{WE}}(w_1, w_2) = \begin{cases} 0 & \text{if } w_1 \text{ or } w_2 \text{ are OOV} \\ v_1 \cdot v_2 & \text{otherwise} \end{cases},$$

where w_1 and w_2 are the words being compared and v_1 and v_2 are their respective embeddings. OOV (Out Of Vocabulary) means that there is no embedding for the word. Despite the large number of embeddings they use, they encounter a lot of OOV terms for n -grams with $n > 1$. To solve this problem, they compose individual word embeddings together, following a simple multiplicative approach where individual vectors of constituent tokens are multiplied together to produce the vector for a n -gram, *i.e.*:

$$W(w) := W(w_1) \times \cdots \times W(w_n),$$

where w is a n -gram composed of n individual word tokens $w_1, \dots, w_n$ and $W(w_i)$ is the embedding of the word w_i . Their experimental results show that ROUGE-WE is able to achieve better correlations with human judgements when measured with the Spearman and Kendall rank coefficients.

BERTSCORE. Zhang et al. [2020] introduce BERTSCORE, a language generation evaluation measure based on pretrained BERT contextual embeddings [Devlin et al., 2019]. BERTSCORE computes the similarity of two sentences as a sum of cosine similarities between their tokens'

embeddings. BERTSCORE addresses two common pitfalls in n -gram-based measures. First, such methods often fail to robustly match paraphrases: this leads to performance underestimation when semantically-correct phrases are penalized because they differ from the surface form of the reference. Second, n -gram models fail to capture distant dependencies and penalize semantically-critical ordering changes. In contrast, contextualized embeddings are trained to effectively capture distant dependencies and ordering. The complete BERTSCORE matches each token in a reference x to a token in a candidate $\hat{x}$ to compute recall, and vice versa to compute precision. Both are combined to compute an F1 measure:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} (\mathbf{x}_i^\top \hat{\mathbf{x}}_j) \quad , \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} (\mathbf{x}_i^\top \hat{\mathbf{x}}_j) \quad , \quad F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} .$$

They additionally use importance weighting and baseline rescaling. They show that BERTSCORE correlates better with human judgments and provides stronger model selection performance than existing measures.

MOVERSCORE. Zhao et al. [2019] introduce MOVERSCORE, a distance metric built upon a combination of (i) contextualized representations of system and reference texts and (ii) a distance between these representations measuring the semantic distance between system outputs and references. They find particularly important for a measure to not only capture the amount of shared content between two texts, *i.e.*, $\text{intersect}(A, B)$, as is the case with many semantic textual similarity measures, but also to accurately reflect to what extent the system text has deviated from the reference, *i.e.*, $\text{union}(A, B) - \text{intersect}(A, B)$, which is the intuition behind using a distance metric. MOVERSCORE generalizes the Word Mover’s Distance proposed by Kusner et al. [2015]: for $\mathbf{x}$ and $\mathbf{y}$ two sentences viewed as sequences of n -grams x_i and y_j and $\mathbf{C}$ a transportation cost matrix such that $\mathbf{C}_{ij} = d(x_i, y_j)$, we have:

$$\text{WMD}(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{F} \in \mathbb{R}^{|\mathbf{x}| \times |\mathbf{y}|}} \langle \mathbf{C}, \mathbf{F} \rangle ,$$

where $\mathbf{F}$ is the transportation flow matrix with $\mathbf{F}_{ij}$ denoting the amount of flow traveling from x_i to y_j , and $\langle \mathbf{C}, \mathbf{F} \rangle$ is the sum of all matrix entries of $\mathbf{C} \odot \mathbf{F}$, where $\odot$ denotes element-wise multiplication. In practice, they use the euclidean distance between the embedding representations of n -grams:

$$d(x_i, y_j) = \|E(x_i) - E(y_j)\|_2 ,$$

where E is the embedding function. Their findings suggest that measures combining contextualized representations with a distance measure perform the best and demonstrate strong generalization capability across tasks.

BARYSCORE. Colombo et al. [2021] introduce BARYSCORE, an evaluation measure which (i) finds the Wasserstein barycentric distributions of contextual encoder layers for a candidate and a reference, and (ii) evaluates these barycentric distributions using the Wasserstein distance. They compare their approach to the similar MOVERSCORE and argue that using an

euclidean distance does not allow a proper evaluation of the geometry induced by the contextualized encoder layers in the Wasserstein space. In other words, MOVERSCORE evaluates a distorted geometry, inducing wrong interpretability of the transportation flow. The advantage of exploiting Wasserstein barycenters over euclidean aggregation relies on rehabilitating this geometry. They show that BARYSCORE achieves higher correlations with human judgment than MOVERSCORE and other BERT-based measures.

DEPTHSCORE. Staerman et al. [2024] introduce DEPTHSCORE, a new discrepancy measure between probability distributions involving the upper-level sets of data depth is introduced. They show that this measure defines a pseudo-metric in general and demonstrate its good behavior regarding major transformation groups, as well as its ability to factor out translations. They also put forth its robustness through the concept of finite sample breakdown point. Finally, they use this pseudo-metric for NLG evaluation and find that DEPTHSCORE achieves higher correlations with human judgment than existing evaluation measures.

S3. Peyrard et al. [2017] propose to learn an evaluation model S3 for evaluating automatic summarization that is based on the human judgements available as part of classical summarization datasets. In particular, they note that any existing automatic scoring measure can be included as features: the model will learn the combination exhibiting the best correlation with human judgments. They use a Support Vector Regression (SVR) to learn their model, using many automatic evaluation measures as features and fitting the regression parameters so that the model is able to predict human judgment scores. They show that S3 performs better than the measures used as its features and has low percentage of failure.

SUMMAQA. Scialom et al. [2019] introduce the SUMMAQA measure for the evaluation of question answering. They compute the average F1 score over each (input, question, answer) triplet to obtain a score called $QA_{F-score}$ and compute the average confidence scores for each triplet to obtain QA_{conf} . Then, they use a Ridge regression to predict the geometric mean of readability and relevance from the two scores. Their best linear model, according to the experiment, was a linear combination of both scores and ROUGE-L:

$$\alpha \text{ ROUGE-L} + \beta QA_{conf} + \delta QA_{F-score}.$$

They improve their model with reinforcement learning and observe that it leads to comparable performance in terms of ROUGE w.r.t. state-of-the-art approaches.

INFOLM. Colombo et al. [2022a] introduce INFOLM, a family of untrained metrics that can be viewed as string-based metrics. INFOLM relies on a pretrained masked language model and makes use of information measures, allowing for its adaptation of various evaluation criteria. Given a reference text $\mathbf{x}$ and a candidate text $\mathbf{y}$, INFOLM recursively masks each token position of both $\mathbf{x}$ and $\mathbf{y}$ to obtain individual masked contexts. Using the language model, it then predicts a probability distribution for each context. The two resulting distributions are averaged into well-formed discrete distributions $p_{\mathbf{x}}$ and $p_{\mathbf{y}}$, and then compared through a

measure of information I (e.g. an α -, γ -, or AB -divergence, an L_p distance or the Fisher-Rao distance). Formally:

$$\text{INFOLM}(\mathbf{x}, \mathbf{y}) := I(p_{\mathbf{x}}, p_{\mathbf{y}}).$$

They demonstrate that INFOLM achieves statistically significant improvements and over 10 points of correlation gains in many configurations on both text summarization and data2text generation.

COVERAGE, DENSITY, and COMPRESSION. Grusky et al. [2018] examine summarization strategies using three measures that capture the degree of text overlap between the summary and article, and the rate of compression of the information conveyed. Given an article text A consisting and the corresponding article summary S , they define the set of extractive fragments $F(A, S)$ as the set of shared sequences of token in A and S . They identify $F(A, S)$ using a greedy process and compute two measures based on it. The COVERAGE measure quantifies the extent to which a summary is derivative of a text. $\text{COVERAGE}(A, S)$ measures the percentage of words in the summary that are part of an extractive fragment with the article:

$$\text{COVERAGE}(A, S) := \frac{1}{|S|} \sum_{f \in F(A, S)} |f|.$$

The DENSITY measure quantifies how well the word sequence of a summary can be described as a series of extractions. They define $\text{DENSITY}(A, S)$ as the average length of the extractive fragment to which each word in the summary belongs. The density formulation is similar to the coverage definition but uses a square of the fragment length:

$$\text{DENSITY}(A, S) := \frac{1}{|S|} \sum_{f \in F(A, S)} |f|^2.$$

Finally, they also use a simple dimension of summarization, the compression ratio, to further characterize summarization strategies. They define COMPRESSION as the word ratio between article and summary:

$$\text{COMPRESSION}(A, S) := \frac{|A|}{|S|}.$$

Summarizing with higher compression is challenging as it requires capturing more precisely the critical aspects of the article text

TEXT LENGTH, NOVELTY, and REPETITION. Fabbri et al. [2021] perform a meta-evaluation of evaluation measures for text summarization. In addition to many existing evaluation measures, they include a few more data statistics:

1. TEXT LENGTH is the length of the sequence;
2. NOVELTY is the percentage of n -grams in the summary not found in the input document;
3. REPETITION is the percentage of n -grams in the summary which are repeated.

SUPERT. Gao et al. [2020] study unsupervised multi-document summarization evaluation measures, which require neither human-written reference summaries nor human annotations. They propose SUPERT, which rates the quality of a summary by measuring its semantic similarity with a pseudo-reference summary, *i.e.*, selected salient sentences from the source documents, using contextualized embeddings and soft token alignment techniques. They show that, compared to the state-of-the-art unsupervised evaluation measures, SUPERT correlates better with human ratings by 18-39%. Furthermore, they use SUPERT as rewards to guide a neural-based reinforcement learning summarizer, yielding favorable performance compared to the state-of-the-art unsupervised summarizers.

BLANC. Vasilyev et al. [2020] present BLANC, a new approach to the automatic estimation of document summary quality. BLANC is defined as a measure of how well a summary helps an independent, pre-trained language model while it performs its language understanding task on a document. It is based on a BERT [Devlin et al., 2019] instance that is pre-trained to predict masked text tokens (words or sub-words). There are two BLANC versions: BLANC-help, which uses the summary text by directly concatenating it to each document sentence during inference, and BLANC-tune, which uses the summary text to fine-tune the language model, and then processes the entire document. BLANC-help is defined as the difference between the accuracy A_s of unmasking with the summary and the accuracy A_f of unmasking with the filler:

$$\text{BLANC}_{\text{help}} := A_s - A_f = \frac{S_{01} - S_{10}}{S_{\text{total}}},$$

where the S_{ij} are the counts of unsuccessful/successful (0/1) unmasking for filler-input/summary-input (i/j). For BLANC-tune, the model first learns from the summary and then they observe how helpful this learning was in reconstructing masked tokens in text sentences by comparing the accuracies of two reconstructions: one that does use the summary, and one that does not. They show that BLANC can be competitive to human annotators for summarization evaluation.

BARTSCORE. Yuan et al. [2021] conceptualize the evaluation of generated text as a text generation problem, modeled using pre-trained sequence-to-sequence models. They posit that models trained to convert generated text to/from a reference output or the source text will achieve higher scores when the generated text is better. They operationalize this idea using BART [Lewis et al., 2020], and propose the BARTSCORE measure, which uses the weighted log-probability of one text $\mathbf{y}$ given another text $\mathbf{x}$. Formally:

$$\text{BARTSCORE} = \sum_{t=1}^m \omega_t \log p(\mathbf{y}_t | \mathbf{y}_{<t}, \mathbf{x}, \theta),$$

where the ω_t are the chosen weights and θ the parameters of the model. They show that BARTSCORE can outperform existing top-scoring measures in 16 of 22 test settings.

Note on the Names of Variants of Evaluation Measures. In this paragraph, we define the names we give to some variants of the automatic evaluation measures we used. SUPERT

and BLANC are summarization measures which normally require a source document and a summary. In the ASG setting, we instead have a prompt and a generated story. The suffix “PS” means that we used the “Prompt as the Summary”, and “SS” means the “Story as the Summary”. The GOLD suffix means we used the reference human story as the source document and the generated story as the summary. Also, given a couple of texts (x, y) , BARTSCORE computes a score based on the log-probability of y given x . We used the suffixes “SH” for “(Story, Human)”, “HS” for “(Human, Story)”, “SP” for “(Story, Prompt)” and “PS” for “(Prompt, Story)”.

3.4 Proposed Framework for Meta-Evaluation

After designing a set of criteria for human evaluation and selecting automatic evaluation measures, we present in this section the framework we will use for our meta-evaluation of ASG.

3.4.1 Related Work

We reviewed the related work for meta-evaluation in general in [Section 2.3.3](#). For ASG specifically, [Guan et al. \[2021b\]](#) introduced the OPENMEVA benchmark which compares human and generated stories. They sample prompts and stories from the ROCSTORIES [[Mostafazadeh et al., 2016](#)] and WRITINGPROMPTS [[Fan et al., 2018](#)] datasets (interpreting the first sentence of the story as the prompt for ROCSTORIES) and generate stories with several models. They also produce negative examples by applying perturbations to human stories. Then, they ask workers to evaluate a batch of stories w.r.t. their “overall quality”. They refine the evaluation through a point deduction policy: if the story contains errors (*e.g.* repetitive plots or unrelated events), its quality is reduced. While their protocol has its merits, we reckon that the stories from ROCSTORIES are too short to be evaluated in a meaningful way (50 words on average), and that the aggregation of all errors in a single criterion is not ideal.

3.4.2 Meta-Evaluation Strategies

Our main objective will consist in comparing evaluation methods between one another. Especially, we would like to ascertain how well automatic measures correlate with human evaluation, which we will use as our gold standard.

Notations. For S systems and N story-prompts, let y_i^j be the story generated by system $j \in \{1, \dots, S\}$ for story-prompt $i \in \{1, \dots, N\}$. For a (human or automatic) measure m , we note $m(y_i^j)$ the score associated to y_i^j . Let K be a correlation coefficient (*e.g.* Kendall’s τ [[Kendall, 1938](#)]).

Definition 7: System-level Correlation ($K_{m_1, m_2}^{\text{sys}}$)

This strategy measures how suited m_1 is w.r.t. m_2 if used to compare the performance of two systems. The correlation is applied to the mean values over all stories for all systems for both measures. Formally:

$$K_{m_1, m_2}^{\text{sys}} := K\left(\frac{1}{N} \mathbf{C}_{m_1}^{\text{sys}}, \frac{1}{N} \mathbf{C}_{m_2}^{\text{sys}}\right), \quad (3.1)$$

where $\mathbf{C}_m^{\text{sys}} := \left[\sum_{i=1}^N m(y_i^1), \dots, \sum_{i=1}^N m(y_i^S) \right]$.

Definition 8: Segment-level Correlation ($K_{m_1, m_2}^{\text{seg}}$)

This strategy measures how suited m_1 is w.r.t. m_2 if used as a loss or reward for a model. The correlation is applied to each story among all system outputs and the mean is taken. Formally:

$$K_{m_1, m_2}^{\text{seg}} := \frac{1}{N} \sum_{i=1}^N K\left(\mathbf{C}_{m_1, i}^{\text{seg}}, \mathbf{C}_{m_2, i}^{\text{seg}}\right), \quad (3.2)$$

where $\mathbf{C}_{m, i}^{\text{seg}} := [m(y_i^1), \dots, m(y_i^S)]$.

However, the segment-level correlation, often used in conjunction with the system-level one in the meta-evaluation literature [Bhandari et al., 2020; Ma et al., 2019], is not adapted to ASE since stories generated from the same story-prompt are not required to be similar, while *e.g.* translations of a sentence should look alike.

We therefore use the overall correlation, which we define below.

Definition 9: Overall Correlation ($K_{m_1, m_2}^{\text{ovl}}$)

This strategy computes the correlation between the full vectors containing the scores of m_1 or m_2 for a given story for every system. Formally:

$$K_{m_1, m_2}^{\text{ovl}} := K(\mathbf{C}_{m_1}, \mathbf{C}_{m_2}), \quad (3.3)$$

where $\mathbf{C}_m := \left[\left(m(y_i^j) \right)_{(i, j) \in \{1, \dots, N\} \times \{1, \dots, S\}} \right]$.

Remark 1: Differences With Our COLING Publication

We introduced the overall correlation strategy in our second publication [Chhun et al., 2024] after reckoning that the segment-level correlation (called “story-level” correlation in Chhun et al. [2022]) was not particularly adapted for the evaluation of ASG, as explained above. We therefore updated the figures and corresponding analyses in this thesis.

Correlation Coefficients. Three correlation coefficients are commonly used for meta-evaluation:

1. Pearson’s r [Pearson, 1895] measures linear correlation between two sets of data X and Y . It is defined as the covariance of the two variables divided by the product of their standard deviations. When applied to a sample, it is formulated as:

$$r_{XY} := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

where n is the sample size, x_i, y_i are the individual sample points and $\bar{x}, \bar{y}$ are the sample means. Pearson’s r is essentially a normalized measurement of the covariance, such that the result always has a value between -1 and 1.

2. Spearman’s ρ [Spearman, 1904] is a nonparametric measure of rank correlation that is defined as the Pearson correlation between the rank values of two variables X and Y ; while Pearson’s correlation assesses linear relationships, Spearman’s correlation assesses monotonic relationships (whether linear or not). Formally:

$$\rho_{XY} := r_{R_X R_Y},$$

where R_X and R_Y are the rank variables associated with X and Y .

3. Kendall’s τ [Kendall, 1938] is a measure of rank correlation that describes the similarity of the orderings of the data when ranked by each of the quantities. It is computed based on the numbers of system pairs out of $\binom{N}{2}$ which are ranked the same between X and Y . Formally:

$$\tau := \frac{P - Q}{\sqrt{(P + Q + T) \cdot (P + Q + U)}},$$

where P and Q are the number of pairs ranked the same or different by X and Z respectively, and T and U are the number of ties only in X or Y respectively.

Since our main concern is to evaluate automatic measures’ ability to correctly determine whether one system is better than another, we favor Spearman’s ρ and Kendall’s τ over Pearson’s r . Moreover, Gilpin [1993] argues that Kendall’s τ approaches a normal distribution more rapidly than Spearman’s ρ as the sample size increases, and that it is more tractable mathematically, particularly when ties are present. Newson [2002] also argues that confidence intervals for Spearman’s ρ are less reliable and less interpretable than confidence intervals for Kendall’s τ . For those reasons, we choose to use Kendall’s τ for our meta-evaluation process.

3.4.3 Statistical Testing

Williams Test. Correlations between two automatic measures on the same annotated dataset are not independent. As advised by [Graham and Baldwin \[2014\]](#), we use the Williams test [[Moon, 2019](#); [Williams, 1959](#)] to evaluate the strength of an *increase* in dependent correlations [[Steiger, 1980](#)]. Given three features X_1 , X_2 and X_3 of a population of size n , Williams’s t test for whether the correlation between X_1 and X_2 equals the correlation between X_1 and X_3 is formulated as follows:

$$t := \frac{(r_{12} - r_{13})\sqrt{(n-1)(1+r_{23})}}{\sqrt{2K\frac{(n-1)}{(n-3)} + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}},$$

where r_{ij} is the correlation between X_i and X_j and

$$K := 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}.$$

Williams’s t statistic follows a Student’s t -distribution with $n - 3$ degrees of freedom. In particular, the Williams test takes the correlations between X_2 and X_3 into account.

p -value Adjustment Procedure. Furthermore, since we perform a large quantity of tests, we choose to correct p -values for multiplicity. A common procedure is to control for the probability of a Type I error, *i.e.*, the probability of erroneously rejecting the null hypothesis, also called the family-wise error rate (FWER). FWER control methods include the Bonferroni correction [[Dunn, 1959](#)], which consists in multiplying the p -values by the number of hypotheses. However, the Bonferroni correction is very conservative, which means that the statistical power is greatly reduced. Thus, [Jafari and Ansari-Pour \[2019\]](#); [Lydersen \[2024\]](#) argue for controlling the false discovery rate: instead of limiting the probability of at least one false discovery, we control for the expected proportion of false discoveries.

We use the Benjamini-Hochberg (BH) method [[Benjamini and Hochberg, 1995](#)]. Given m p -values $p_1, \dots, p_m$ sorted in increasing order and a significance level α , the Benjamini-Hochberg method consists in finding the largest k such that $p_k \leq \frac{k}{m}\alpha$. The null hypothesis would then be rejected for the first k tests. This is equivalent to computing adjusted p -values $p_k^* = p_k \frac{m}{k}$ and replacing the p -values from largest to smallest.

Gradual Notion of Evidence. Following recent recommendations to move beyond simplistic “statistical significance” tests [[Amrhein et al., 2019](#); [McShane et al., 2019](#); [Wasserstein et al., 2019](#)], we report all p -values for transparency. So as to be less dichotomous about the interpretation of statistical results, we choose to use a gradual notion of evidence for our statistical analysis, as suggested by [Muff et al. \[2022\]](#). The proposed terminology is shown on [Table 3.4](#).

Value Range	Terminology
$1 < p < 0.1$	Little or no evidence
$0.1 < p < 0.05$	Weak evidence
$0.05 < p < 0.01$	Moderate evidence
$0.01 < p < 0.001$	Strong evidence
$0.001 < p < 0.0001$	Very strong evidence

Table 3.4: Gradual evidence language proposal [Muff et al., 2022].

3.5 Conclusion

In this chapter, we presented a new, extensive methodology specifically designed for conducting a meta-evaluation of Automatic Story Generation. First, in order to bridge the gap between the non-standardized human evaluation observed in the NLP literature and the mostly theoretical frameworks of the social sciences literature, we proposed an original, comprehensive set of six criteria for human story evaluation. Then, to find out what type of automatic measures is the most appropriate for Automatic Story Evaluation, we reviewed a wide variety of them, introduced a new two-dimensional taxonomy relevant to ASG, notably through the reference-based *vs.* reference-free distinction, and selected measures from each of our taxonomy’s categories. Finally, we adapted the meta-evaluation framework used in the NLP literature to ASG: in particular, we substitute the less relevant segment-level correlation strategy for the overall correlation strategy, and justify our choice of Kendall’s τ as well as our statistical testing procedure. In [Chapter 4](#), we recount the building process of HANNA, our corpus of Human ANnotated NArratives for ASG evaluation, which allows us to assess whether our proposed human criteria are well-founded. To that end, we will perform an annotation campaign where human raters are asked to grade stories w.r.t. our six criteria ([Section 4.3.1](#)). We will also perform another annotation campaign with a different protocol, which involves Large Language Models ([Section 4.4.2](#)).

Chapter 4

HANNA: A Corpus of Human-ANnotated Narratives for ASG Evaluation

*“ This is a story of the past. The story
of him and her, only fifteen days
long, concerning a certain grail. ”*

Fate/stay night: Last Episode

Contents

4.1 Introduction	64
4.2 Design Decisions for Building HANNA	65
4.2.1 Survey of Corpora Related to Automatic Story Generation	65
4.2.2 State-of-the-Art Models for Automatic Story Generation	69
4.3 The HANNA Corpus	74
4.3.1 Building a First Version of HANNA	74
4.3.2 Inter-Rater Reliability	79
4.3.3 Evaluating Our Human Criteria	80
4.3.4 Comparing ASG Systems	81
4.3.5 Need for a HANNA Upgrade	82
4.4 Augmenting HANNA with Large Language Models	82
4.4.1 Experimental Protocol for ASG using LLMs	84
4.4.2 Experimental Protocol for Automatic Annotation using LLMs	85
4.4.3 LLM Performance at ASG	92
4.5 Conclusion	92

Chapter 4 Abstract

In order to use our human evaluation criteria (Section 3.2) and to conduct a meta-evaluation with our proposed framework (Section 3.4), we need an appropriate dataset of stories. In this chapter, we describe the building process of HANNA, our corpus of Human ANnotated NArratives. We begin by reviewing corpora related to Automatic Story Generation and find that no existing dataset contains human annotations w.r.t. specific criteria of story quality (Section 4.2.1). We then review and select ASG systems (Section 4.2.2) in order to build our own corpus, HANNA, specifically designed for ASG evaluation. The first version of HANNA contains 1,056 stories produced by 11 systems and aligned on 96 prompts (Section 4.3.1). We perform an annotation experiment on HANNA V1, asking human raters to grade each story w.r.t. our six human criteria. We analyze the overall and system-level correlations between our human criteria, and confirm that they allow for a standardized and extensive human evaluation (Section 4.3.3). We also observe that large pre-trained language models, namely GPT-2, produce the best results for ASG (Section 4.3.4). Finally, we use Large Language Models to augment HANNA with 480 new stories (Section 4.4.1) and 150k+ annotations (Section 4.4.2), and observe that LLMs are rated more highly than humans when evaluated by two different LLMs (Section 4.4.3).

4.1 Introduction

Performing evaluation experiments in NLP requires appropriate resources and, especially, a fitting dataset. NLP research in particular has a long history of developing corpora and benchmarks to support the development and evaluation of algorithms for multiple tasks.

Many early initiatives focused on large-scale annotation collection to assist data-driven methods for low-level tasks like part-of-speech tagging [Marcus et al., 1993] and named entity recognition [Grishman and Sundheim, 1996]. Other early benchmarks were primarily intended to explore the information found in linguistic knowledge and context, such as the Message Understanding Conference (MUC) benchmarks for information extraction [Sundheim, 1993] and the Automatic Content Extraction (ACE) benchmarks for relation extraction and coreference resolution [Doddington et al., 2004]. Other benchmarks were centered on reading comprehension [Hirschman et al., 1999] and question answering [Kočíský et al., 2018].

The scope of these benchmarks has ranged from more general tasks such as textual entailment [Dagan et al., 2006] to very specialized tasks like reference resolution [Levesque et al., 2012]. Some benchmarks, like GLUE [Wang et al., 2019], are made up of multiple tasks, whereas others concentrate primarily on one task only: for instance, the SQUAD benchmark [Rajpurkar et al., 2016] contains more than 100,000 questions for machine comprehension.

In this chapter, we elaborate on the building process of HANNA, our corpus of Human ANnotated NArratives designed for the evaluation of Automatic Story Generation.

4.2 Design Decisions for Building HANNA

We begin by surveying existing corpora of written stories (Section 4.2.1) to examine whether how appropriate they are for our benchmarking goals; in particular, we look for corpora with extensive human annotations w.r.t. specific evaluation criteria.

4.2.1 Survey of Corpora Related to Automatic Story Generation

ROCSTORIES. The first major, easily accessible dataset of stories found in the recent literature is ROCSTORIES [Mostafazadeh et al., 2016], a crowdsourced corpus of 50k 5-sentence stories with titles. ROCSTORIES was designed for the Story Cloze Test, an evaluation framework for story understanding that consists in predicting the final sentence of a story given the four previous sentences. ROCSTORIES includes diverse causal and temporal relationships across a broad spectrum of situations, and its multiple nonfictional short stories constitute effective training data for building story generation models. Examples of stories from ROCSTORIES can be found in Table 4.1. As part of the Story Cloze Test evaluation framework they introduce, every story has an alternative “wrong” ending that complements the canonical “right” ending (see Table 4.2).

Title	Five-sentence Story
The Test	Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week. Jennifer felt bittersweet about it.
The Hurricane	Morgan and her family lived in Florida. They heard a hurricane was coming. They decided to evacuate to a relative's house. They arrived and learned from the news that it was a terrible storm. They felt lucky they had evacuated when they did.
Spaghetti Sauce	Tina made spaghetti for her boyfriend. It took a lot of work, but she was very proud. Her boyfriend ate the whole plate and said it was good. Tina tried it herself, and realized it was disgusting. She was touched that he pretended it was good to spare her feelings.

Table 4.1: Example stories from ROCSTORIES [Mostafazadeh et al., 2016].

SIND. Huang et al. [2016] developed SIND (Sequential Images Narratives Dataset), the first dataset for sequential vision-to-language, for the task of visual storytelling. SIND contains 81,743 unique photos in 20,211 sequences, with accompanying captions and stories. The data is organized in three categories of language, so as to assess the importance of temporal context and narrative language: (1) Descriptions of images-in-isolation (DII), (2) Descriptions of images-in-sequence (DIS) and (3) Stories for images-in-sequence (SIS). They train RNN-based story generation models on SIND and, using simple heuristics, perform experiments

Context	Right Ending	Wrong Ending
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a \$10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She found her phone in the car.	She didn't want her phone anymore.

Table 4.2: Example stories from the Story Cloze Test [Mostafazadeh et al., 2016].

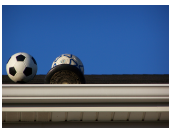
	1	2	3	4	5
					
DII	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
DIS	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up.	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
SIS	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

Table 4.3: Example story from SIND [Huang et al., 2016].

that demonstrate that story generation is fundamentally different from caption generation. Examples are shown on Table 4.3.

Prompt	Story
You’ve finally managed to discover the secret to immortality. Suddenly, Death appears before you, hands you a business card, and says, “When you realize living forever sucks, call this number, I’ve got a job offer for you.”	So many times have I walked on ruins, the remainings of places that I loved and got used to.. At first I was scared, each time I could feel my city, my current generation collapse, break into the black hole that thrives within it, I could feel humanity, the way I’m able to feel my body.. After a few hundred years, the pattern became obvious, no longer the war and damage that would devastate me over and over again in the far past was effecting me so dominantly. [...]
The moon is actually a giant egg, and it has just started to hatch.	- Week 18 aboard the Depth Reaver, Circa 2023 - I walk about the dull gray halls, the artificial gravity making my steps feel almost as if they were on land. Almost. I glance out a window as I pass it by. There’s the sun, and there’s the moon right there. And, of course, there’s the Earth. I kinda miss it. Then again, space is pretty cool. It’s got some brilliant views, and the wifi is surprisingly good. [...]
Write a horror story from the perspective of the antagonist. Make them as sympathetic as possible.	I don’t want to cut off his head, but I don’t really have a choice. I close my eyes and just wait for it to be over. My insides turn as I feel the swing connect. <i>Quit being so weak! You’ve done this before.</i> The jock’s girlfriend screams on cue, looks like she is going to faint. She has blonde hair and a blue dress patterned with diamonds. I see how scared she is, and I feel ashamed. I’m not the bad guy. [...]

Table 4.4: Example stories from WRITINGPROMPTS [Fan et al., 2018].

WRITINGPROMPTS. Fan et al. [2018] released the WRITINGPROMPTS dataset, which contains 300k human-written stories paired with writing prompts from Reddit’s /r/WritingPrompts forum, where members encourage one another to write stories by providing story premises called prompts. Those prompts exhibit a wide range of themes and can be answered by multiple users. To be included in the dataset, stories had to be longer than 30 words, avoid inappropriate content, and be related to the prompt. They also introduce a hierarchical model with self-attention for story generation, and demonstrate that it produces stories which are more fluent and more relevant to the prompt than previous state-of-the-art models. Examples are shown on Table 4.4.

ROLEPLAYERGUILD. Louis and Sutton [2018] collect a corpus of RPG threads from the roleplayerguild.com website. RPGs are usually played in-person, but users can also play them online. Threads come in pairs: the first thread contains the description of the players’ characters, with multiple attributes, and the second thread contains the actual game play where each player describes the actions they take when their turn comes. The ROLEPLAYERGUILD corpus is a collection of 1,544 RPGs of different themes, such as adventure, romance and horror. An example is shown on Table 4.5.

Character Description	Action Description
Name: Ana Blackclaw; Age: 27; Gender: Female Appearance: Standing at a mighty 6'5, she is a giant among her fellow humans. Her face is light, though paler than the average man or woman's, and is marked by scars. ... Her body is muscular, as it would have to be to carry both her armor and the hammer. Her light grey eyes nearly always keep a bored expression. Her canines seem a tad larger than the normal person's. Preferred Weapon: Hammer. Preferred Armor: Heavy. Gift: Binoculars. Darksign: No.	She stopped dead in her tracks as the hissing began. A grumble escaped her as it did so, and she looked over to make sure the other woman was doing fine. Seeing that all was not entirely well, she allowed herself to slide down, her hand gripping the slope side once more to slow herself. Once that was accomplished, she reached out and grabbed the back of the girl's neck, pulling her back to steady herself. The giant remained silent as she did so, and then glanced over to the nearby skeletons. They would be upon them soon. Her grip tightened on the hammer as she glanced from side to side. It would not be a fun fight.

Table 4.5: Example descriptions from ROLEPLAYERGUILD [Louis and Sutton, 2018].

PG-19. Rae et al. [2020] propose the PG-19 dataset, a new open-vocabulary language modelling benchmark derived from Project Gutenberg books that were published before 1919. The PG-19 dataset contains 28,752 books, which they use for evaluating their Compressive Transformer model against other state-of-the-art models. They find that training on PG-19 helps their model produce long-range stories with different writing styles.

STORIUM. Akoury et al. [2020] release a dataset of 6k lengthy stories (125M tokens) collected from the STORIUM collaborative storytelling community. Each story comes with multiple human annotations describing the characters' background and attributes. Whereas STORIUM contains fewer stories than other datasets such as ROCSTORIES and WRITINGPROMPTS, its stories are overall longer and contain on average 14 narrator prompts per story and 41 annotations. They integrate language models onto STORIUM to enable users to generate and edit story continuations. They compute automatic evaluation measures scores over those edits and find that they correlate well with human ratings and feedback from user studies.

OPENMEVA. Guan et al. [2021b] propose OPENMEVA, a benchmark for evaluating open-ended story generation metrics. OpenMEVA can be used to perform multiple tests on evaluation measures: (a) correlation with human judgments; (b) generalization to various model outputs and datasets; (c) story coherence judgment; and (d) robustness to perturbations. OPENMEVA comes with both automatically-generated test cases and manually-annotated stories. They evaluate current measures on OPENMEVA and find that they fail to recognize incoherence at the discourse-level correlate poorly with human judgments.

Name	Type	Annotations	# Words
ROCSTORIES	Title + Story	✗	80
SIND	Pictures + Story	✗	80
WRITINGPROMPTS	Prompt + Story	✗	750
ROLEPLAYERGUILD	RPG Thread	✗	3,000
PG-19	Book	✗	69,000
STORIUM	Collaborative Story	≈	19,000
OPENMEVA	Title/Prompt + Story	≈	400

Table 4.6: Overview of existing story generation corpora. No corpus provides annotations on different criteria of story quality.

■ Takeaways

An overview of the reviewed corpora can be found in Table 4.6. Most notably, none of the datasets we reviewed provided annotations w.r.t. specific evaluation criteria for story quality. STORIUM contains annotations that indicated whether story components were relevant to one another, while the OPENMEVA dataset [Guan et al., 2021b] contained human annotations on a single criterion, “overall quality”, which is negatively defined w.r.t. several error types. We therefore set out to create our own dataset for ASG evaluation. To that end, we need to select story generation models: we review the existing ASG systems in the next section.

4.2.2 State-of-the-Art Models for Automatic Story Generation

The Automatic Story Generation task has been tackled with a variety of approaches, as shown in Section 2.1. However, neural network-based models now constitute the vast majority of ASG models: we review a number of those below. We can make a distinction between older RNN-based approaches and newer transformer-based approaches.

■ RNN-based ASG Systems

DEEPTINGLE. One of the first neural network-based ASG models is DEEPTINGLE, introduced by Khalifa et al. [2017]. DEEPTINGLE is made of two main components: (1) the neural network that learns to generate words in the style of the training corpus and (2) a set of co-creativity tools which help refine the writing style. DEEPTINGLE’s training data contains all the books by Chuck Tingle released until November 2016, for a total of 109 short stories and 2 novels, all with similar structure. DEEPTINGLE’s neural network is a 6-layer architecture based on LSTMs and GLOVE embeddings. Through a user study, they show that DEEPTINGLE is better than a Markov model at generating coherent and fluent text.

EVENT2EVENT and EVENT2SEQUENCE. Martin et al. [2018] present a technique for translating textual story data into event sequences. They decompose the task of ASG into the

generation of successive, generalized events (EVENT2EVENT) and the generation of natural language sentences from events (EVENT2SEQUENCE). Both EVENT2EVENT and EVENT2SEQUENCE networks are recurrent multi-layer RNN encoder-decoder networks. First, a human-written sentence is translated as one or more events, and the EVENT2EVENT network generates successive events. The EVENT2SEQUENCE network then translates the events back into natural language. They show that generalized events with full-length generalized sentences yields an increase BLEU scores.

ENGEN. Clark et al. [2018] propose an entity-based generation model which uses three sources of contextual information: the current sentence, the previous sentence, and the current state of all mentioned entities. ENGEN combines S2SA, an LSTM-based entity-unaware model with an attention mechanism, with ENTITYNLM [Ji et al., 2017], a language model that adds explicit tracking of entities. They show that the combination of the three contexts improves model performance on three tasks: mention generation, sentence selection, and sentence generation.

Controllable Story Generation. Peng et al. [2018] propose a general framework for interactive story generation focused on the extraction of two control factors, ending valence and storyline, and two generation tasks. An ending valence is a label that describes a happy, sad, or unknown ending, while a storyline is a sequence of words that summarizes the story. Both control factors are extracted from the ROCSTORIES dataset. The task of ending-valence controlled generation consists in the generation of a story ending given the first four sentences of a story and an ending valence. They use an LSTM-based conditional language model for it. For storyline-controlled generation, the goal is to generate a five-sentence story from a storyline. Here, they use a SEQ2SEQ model with attention. Through automatic and human evaluation, they show that their models generate stories that are both more coherent and more faithful to human inputs.

FUSION. Fan et al. [2018] introduce a language model capable of hierarchical story generation, *i.e.*, that is able to decompose the generation process into two levels. First, they use a convolutional language model [Dauphin et al., 2017] for the generation of a story premise called a prompt. Then, the prompt is fed to a SEQ2SEQ model tasked to generate a relevant story. Conditioning on the prompt makes the story more consistent and better-structured. For this, they use a convolutional SEQ2SEQ model [Gehring et al., 2017], based on deep convolutional networks for both the encoder and decoder; they also supplement the decoder with a self-attention mechanism to enable modeling of unbounded context. Finally, they adopt a fusion-based approach using a cold fusion mechanism [Sriram et al., 2018]: they train a SEQ2SEQ model that has access to the hidden states of a pretrained SEQ2SEQ model. This process is similar to residual learning and enables the second model to focus on conditioning on the prompt, since the first model was not trained to do that. They show through automatic and human evaluation that their proposed model produces stories that are more fluent and more relevant to the prompt than previous state-of-the-art.

Skeleton-Based Model. Xu et al. [2018] propose a skeleton-based model focused on generating coherent stories. While most generation models produce stories in a linear, auto-regressive manner, their model first generates a skeleton that contains the most important words of the future sentence, and then expands the skeleton to a complete sentence. The skeleton generation process is learned through a reinforcement learning method. Through human and automatic evaluation, their skeleton-based model is shown to generate more coherent text than existing state-of-the-art models.

Event-to-Sentence. Ammanabrolu et al. [2020] observe that SEQ2SEQ models often ignore the input event when generating a sentence, and look for methods neural language generation capable of preserving the semantic meaning of the input. They propose an ensemble-based model that relies on four different sub-models and a baseline fifth sub-model that is used as a fallthrough. Each sub-model is tasked to translate events into “generalized” sentences, where nouns are replaced by WordNet Synsets. If a sub-model does not produce a satisfying result, the system continues onto the next sub-model. They demonstrate that each sub-model in the ensemble is calibrated toward a different objective, which enables them to cover for one another’s weaknesses. Through human and automatic evaluation, they show that their model generates more coherent and believable stories than other baseline approaches.

PLAN-AND-WRITE. Yao et al. [2019] propose the PLAN-AND-WRITE hierarchical generation framework, that first plans a storyline, and then generates a corresponding story. They compare two approaches to planning: the *static* schema maps out the complete plot before generating stories, and the *dynamic* schema integrates story planning with its surface realization in text. Both schemas are based on bidirectional RNN language models. Experiments on ROCSTORIES [Mostafazadeh et al., 2016] show that both schemas outperform the baselines, and that human evaluators preferred stories generated by the static schema over the dynamic one.

Coarse-to-Fine Strategies. Fan et al. [2019] introduce a model capable of decomposing the story generation process into a series of easier problems. Initially, the model creates the text’s predicate-argument structure, labeling distinct references to the same entity using placeholder tokens. After that, it creates a surface realization of the predicate-argument structure and, in the end, replaces the placeholders with context-sensitive names. The model is based on a convolutional SEQ2SEQ architecture [Gehring et al., 2017]. In order to enhance the model’s ability to represent long-range context, the decoder uses a gated multi-head self-attention mechanism [Fan et al., 2018]. They discover that the stories produced by their model are preferred by human judges over a variety of other generation models.

Latent Discrete Planning. Jhamtani and Berg-Kirkpatrick [2020] propose a deep latent variable model that samples a sequence of anchor words before generating a sentence for each of them. Its decoder architecture follows from the PLAN-AND-WRITE model [Yao et al., 2019]. During training, the model represents the sequence of anchor words as a latent variable and learns to generate sequences that are effective at guiding the generation process. Through

human evaluation, they demonstrate that their model produces stories that are rated more highly than those of other baselines.

■ Transformer-based ASG Systems

Targeted Commonsense Grounding. Mao et al. [2019] consider the challenge of common sense reasoning (CSR) in language modeling for ASG. They first perform an intermediate fine-tuning [Howard and Ruder, 2018; Phang et al., 2018] of GPT-2 on the BOOKCORPUS dataset [Zhu et al., 2015]. Next, they fine-tune the model with a multi-task learning objective: the training consists in both a language modeling task on WRITINGPROMPTS and a perplexity ranking on both the SWAG dataset [Zellers et al., 2018] and a custom-made dataset containing both human and machine-generated samples. They show that their model achieves higher perplexity on the WRITINGPROMPTS dataset than other baselines.

Narrative Interpolation. Wang et al. [2020] propose a method for controlled story generation aimed at producing stories with user-specified endings. Their GPT-2-based interpolation model fills in the gaps in a narrative by conditioning on both the beginning and the ending of the story. A re-ranker based on ROBERTa also aids in maintaining the consistency of the output content. They conduct a human evaluation and show that, compared to previous methods, ending-guided generation produces more coherent stories and requires less human supervision.

Emotion-Aware Storytelling. Brahman and Chaturvedi [2020] propose narrative techniques for generating stories that follow predetermined story titles and protagonist emotion arcs. Their three models are based on GPT-2: the Emotion Supervision model embeds the emotion arc as an additional input, while the Emotion-Reinforced models make use of unique rewards that are either based on an emotion classifier or a commonsense transformer. These rewards are intended to leverage reinforcement learning to regularize the story generation process. Through both human and automatic evaluations, they find that these models greatly outperform baseline approaches.

PLOTMACHINES. Rashkin et al. [2020] propose the task of *outline-conditioned story generation*, which involves creating a coherent narrative w.r.t. a provided outline. The outline is defined as a list of the story’s main characters and events. They then present PLOTMACHINES, a neural narrative model built on top of GPT [Radford et al., 2018]. PLOTMACHINES learns to track the dynamic plot states of a story in a memory matrix in order to translate an outline into a coherent story. In addition, they label training data with high-level discourse information so that the model can learn to adjust its writing styles depending on the stage of the story. Through automatic evaluation, they show that PLOTMACHINES produces more cohesive narratives than other state-of-the-art models.

Knowledge-Enhanced Pretrained Model. Guan et al. [2020] propose a knowledge-enhanced pretrained model based on GPT-2 [Radford et al., 2019] for commonsense story generation.

They post-train their model on knowledge examples built from external knowledge bases: CONCEPTNET 5 [Speer and Havasi, 2012] and ATOMIC [Sap et al., 2019]). To improve the model’s capacity to handle causal and temporal dependencies between sentences, they add a classification objective to the model, which is asked to distinguish true and fake stories during fine-tuning. They find, through automatic and manual evaluation, that their model can generate more stories that are more fluent and coherent than state-of-the-art baselines.

Content Planning with Aristotelian Rescoring. Goldfarb-Tarrant et al. [2020] posit that high-quality content planning solves several issues that are inherent to story generation, and propose a model that learns to create well-structured stories. They utilize a plot-generation language model based on BART [Lewis et al., 2020], which is combined with multiple rescoring models based on ROBERTA. Each rescoring model is trained on a classification task focused on a particular concept from Aristotle’s *Poetics*. They find that their model produces stories that are both more relevant to the prompt and more fluent compared to other baselines, according to both automatic and human evaluation.

Conditional Variational Autoencoder. Fang et al. [2021] propose a transformer-based conditional variational autoencoder (CVAE) for story generation. The model is made of an encoder, a decoder and a variational posterior, which are all built on top of GPT-2 [Radford et al., 2019]. However, unlike usual transformer-based models that learn a sequence of embedding, their CVAE model is dedicated to learning a single vector as an explicit latent representation. They show that their model obtains higher evaluations from automatic measures than previous state-of-the-art models.

HINT. Guan et al. [2021a] observe that generation models such as BART [Lewis et al., 2020] struggle to maintain a coherent event sequence throughout the story. They speculate that this is due to the decoder’s inability to capture discourse structures and high-level semantics at a higher level than simple token-level co-occurrence. They propose HINT, a long text generation model based on BART where the decoder is able to represent the context information at both sentence level and discourse level through the use of special tokens. They propose two pretraining objectives: (1) semantic similarity prediction, and (2) sentence order discrimination. Through automatic and human evaluation, they show that HINT can generate more coherent texts than state-of-the-art baselines.

End-of-Paragraph and End-of-Sequence Tokens. Bai et al. [2021] study how the implicit “unwritten” information contained in end-of-paragraph (EOP) and end-of-sequence (EOS) tokens affects the quality of text generation. Specifically, they conduct experiments with GPT-2 on a story generation task involving fine-tuning with or without special tokens. They show that learning to generate correct EOP tokens produce better stories w.r.t. both human and automatic evaluation.

TD-VAE. Wilmot and Keller [2021] propose a latent vector planning approach based on a Temporal Difference Variational Autoencoder (TD-VAE), which is used for both conditioning

and reranking the story generation process. They build a hierarchical model that stacks, on top of GPT-2, a transformer-based sentence encoder trained to build rich sentence representations and a TD-VAE module that learns to infer and reconstruct future sentence embedding states. They find that human evaluators rated stories generated with TD-VAE reranking with similar or higher grades compared to other models.

KEYWORD2TEXT. Pascual et al. [2021] present a plug-and-play decoding method for controlled language generation. Their main idea is to shift the probability distribution over the vocabulary towards words that are similar to a given topic or keyword. They test their model-agnostic method on GPT-2, mainly studying the influence of hyperparameters. Through two user studies, they find that their method outperforms competing baselines and can be used to enforce the apparition of guide words without compromising the quality of the generated text.

4.3 The HANNA Corpus

4.3.1 Building a First Version of HANNA

In this section, we describe the first version of our HANNA corpus, which we release on GitHub¹ and will be henceforth referred to as “HANNA V1”. A second version of HANNA is augmented with stories and annotations generated by LLMs and will be presented in Section 4.4.

WRITINGPROMPTS. From all the datasets we described in Section 4.2.1, the two most widely used in the literature are ROCSTORIES and WRITINGPROMPTS. Since no existing corpus contained human annotations w.r.t. specific criteria, we decide to choose one of those two datasets as a basis for our own corpus that we would use for our benchmark. While ROCSTORIES has been used in several ASG works [Brahman et al., 2020; Guan et al., 2020, 2021a; Jhamtani and Berg-Kirkpatrick, 2020; Mao et al., 2019; Pascual et al., 2021; Peng et al., 2018; Yao et al., 2019], the shortness of the stories makes it poorly suited for benchmarking the more recent ASG models, which are able to generate stories much longer than five sentences. Thus, we choose to build our corpus as an extension of the WRITINGPROMPTS dataset, motivated both by the fact that it has also been extensively used in previous literature for the design of ASG models [Bai et al., 2021; Fang et al., 2021; Goldfarb-Tarrant et al., 2020; Guan et al., 2021a; Mao et al., 2019; Rashkin et al., 2020; Wilmot and Keller, 2021], and by the fact that the human stories it contains are longer and written by human authors who are members of /r/WritingPrompts, an online community centered around writing.

ASG-Specific Models. In order to perform our benchmark of existing ASG models, we directly contact the authors of articles that introduced ASG systems and ask for the output stories of their systems. We manage to collect the outputs of 5 ASG systems: FUSION [Fan et al., 2018], the Aristotelian Rescoring-based model from [Goldfarb-Tarrant et al., 2020], HINT [Guan et al.,

¹<https://github.com/dig-team/hanna-benchmark-asg/tree/coling>

2021a], the EOP-based model from [Bai et al., 2021], and TD-VAE [Wilmot and Keller, 2021]. We align the outputs we collect to find which are generated from common prompts of the WRITINGPROMPTS dataset: we are able to extract the aligned output stories from the following **3 ASG-specific systems**:

1. FUSION [Fan et al., 2018],
2. HINT [Guan et al., 2021a],
3. TD-VAE [Wilmot and Keller, 2021],

which are presented in greater detail in Section 4.2.2. In total, we extract 96 stories aligned on 96 common prompts. See Table 4.7 for an example of aligned stories. The 96 prompts can be found in Section A.1.

Story-prompt		
A soldier on the front dies in the middle of writing a letter home. It is finished and sent by the man who killed him.		
FUSION Story	HINT Story	TD-VAE Story
“This place is so clean. I can’t see anything. I can’t see anything.” I said to my friend. “I can’t go to the shelter anymore.” he said, his voice dripping sweat and sweat. He looked down at the ground, and took a deep breath. “I don’t want to go to the shelter.” He looked me in the eyes. They were filled with anger, anger and fear. [...]	the messenger took his time reading the letter. the last letter, the one that was never written. the one that had been written, the one that had been written on his face. the one that had been written on his face. the one that had been written on his face. the one that had been written on his face. the one that had been written on his face. [...]	’writes the woman once more in her head. ’Thumbs up, she mumbles as he looks out the window, at the beautiful sunset that is getting darker every day. ’At this point she notices something strange: her letters have gone grey with brown ink. ’I must write to you as soon as possible, Mrs. Perry. ’The soldier raises an eyebrow. [...]

Table 4.7: Example aligned stories from FUSION, HINT, and TD-VAE.

Pretrained Language Models. We then fine-tune **7 pretrained language models** for ASG on a causal language modeling task on WRITINGPROMPTS to generate stories on the same 96 prompts, using the TRANSFORMERS library [Wolf et al., 2020]². We train the following language models:

1. BERTGENERATION (BERTGEN) [Rothe et al., 2020],
2. CTRL [Keskar et al., 2019],
3. ROBERTA [Liu et al., 2019],
4. XLNET [Yang et al., 2019],
5. GPT [Radford et al., 2018],
6. GPT-2 [Radford et al., 2019],

²<https://huggingface.co/transformers>

7. GPT-2 (TAG), an instance of GPT-2 that we trained with <EOP> (“End Of Prompt”) tags, as inspired by Bai et al. [2021], who find that such tags can improve generation performance.

Experimental Protocol for Human Annotations In order to evaluate the HANNA stories w.r.t. our six criteria, we conduct an annotation campaign on Amazon Mechanical Turk. The campaign was performed on HANNA V1, as the V2 did not exist yet.

- As advised by Karpinska et al. [2021], for each task, we provide the human story alongside the story to be annotated, so that the workers could calibrate their judgment.
- Each of the stories is evaluated by three workers on the six proposed criteria.
- For the evaluation, we choose a 5-point Likert scale rather than a rank-based comparison because we reckon that it would be tedious for the annotators to order the 11 evaluated systems.
- We estimate that a HIT should take between 90 and 120 seconds, so we set the remuneration at \$0.28 per HIT, or between \$8.40 and \$11.40 per hour.
- To ensure that annotators spoke fluent English, we restrict access to the experiment to workers located in the UK, the US, Canada, Australia and New Zealand.
- We further require them to have the Masters Qualification.
- To remove noisy annotations and ensure that the workers read the stories, we choose to reject judgments that were made in fewer than 30 seconds.
- Additionally, to ensure that workers did read the story, we ask them to write down the name of the first-mentioned fictional character of the story.
- Finally, following the recommendations of Shapira et al. [2019], we obtain the final annotations of a story by averaging the ratings of the three workers.

The detailed instructions of the experiment along with the detailed rating guidelines can be found in Table 4.8. Examples of annotations are shown in Table 4.9. In total, we collect 19,008 human annotations.

Amazon Mechanical Turk Example Task

Please read the prompt, the human story and the subject story (both stories might be the same). The story you will have to rate is the **subject story**.

Important: we will reject HITs which were done in **fewer than 30 seconds** (unless both stories are exceptionally short). Please rest assured: if you take the work seriously, we have no reason to reject it.

Disclaimer: some stories have been automatically generated and might contain explicit or offensive content.

Note: some stories have been abruptly cut in the middle of a sentence. Please rate them as if they ended just before the unfinished sentence.

Note: if the story is not relevant with respect to the prompt, it **only** affects the **Relevance** criterion! Do not rate 1 everywhere, or we will **reject**!

Then, please write down the name of the **first character** that is mentioned in the **subject story**; if no name is mentioned, write “None”. Only proper nouns count as names.

Then, please rate the **subject story** on a scale from 1 (worst) to 5 (best) on the following criteria: relevance, coherence, empathy, surprise, engagement, and complexity.

Prompt — When you die the afterlife is an arena where you face every insect and animal you killed in your life. If you win you go to heaven, lose you go to hell. Your job was an exterminator on earth.

Human story — 3,000 years have I been fighting. Every morning, the raccoons scratch at my eyes. Every evening, the skunks spray me while the opossums chew at my feet. I have never had any tools. I have only my hands. I don't remember the place I came from before this. All I remember is the daily fight between me and these animals. [...]

Subject story — First of all, not everyone was entitled to be an exterminator. But the ones that were – maybe were, like, *genius*, because, yes, I had once belonged to a less fortunate class of people – had all the opportunity to work for the damn plant killer, and it's hard work. And the horrifying truth is, once you die, and the entire planet turns into a glade [...]

Name of the first character mentioned in the subject story — [Area to fill]

Relevance (measures how well the story matches its prompt):

- 1 — The story has no relationship with the prompt at all.
 - 2 — The story only has a weak relationship with the prompt.
 - 3 — The story roughly matches the prompt.
 - 4 — The story matches the prompt, except for one or two small aspects.
 - 5 — The story matches the prompt exactly.
-

Coherence (measures whether the story makes sense):

- 1 — The story does not make sense at all. For instance, the setting and/or characters keep changing, and/or there is no understandable plot.
- 2 — Most of the story does not make sense.

3 — The story mostly makes sense but has some incoherences.
4 — The story almost makes sense overall, except for one or two small incoherences.
5 — The story makes sense from beginning to end.

Empathy (measures how well you understood the characters' emotions, regardless of whether you agreed with them):
1 — The characters seemed apathetic to you.
2 — At least one character slightly related to you on an emotional level.
3 — You recognized specific, but not necessarily strong, emotions (<i>e.g.</i> sadness, joy, fear. . .) in at least one character.
4 — At least one character emotionally involved you, but minor details prevented you from completely relating to them.
5 — At least one character completely involved you on an emotional level.

Surprise (measures how surprising the end of the story was):
1 — The ending seemed completely obvious from the start, or doesn't make any sense at all.
2 — The ending was easily predictable after a few sentences.
3 — The ending was predictable after half of the story.
4 — The ending surprised you, but would have been difficult to predict.
5 — The ending surprised you, and still seemed as if it could very reasonably have been predicted, <i>i.e.</i> , there were enough clues in the story.

Engagement (measures how much you engaged with the story):
1 — You found the story boring and were glad it was over.
2 — You found one or two things interesting in the story, but no more.
3 — The story was mildly interesting.
4 — The story almost kept you engaged until the end.
5 — You were so engaged that you wished there was a sequel.

Complexity (measures how elaborate the story is):
1 — The setting of the story is extremely simple; it only involves one or two characters or concepts.
2 — The setting of the story is simple; one or two characters, a simple plot, maybe an indication of time or location.
3 — The story is somewhat developed: it involves at least one of the following: complex concepts, realistic characters, an intricate plot, an underlying history or circumstances, precise descriptions.
4 — The story is developed: it involves at least two of the following: complex concepts, realistic characters, an intricate plot, an underlying history or circumstances, precise descriptions.
5 — The story is well thought-out: it involves at least three of the following: complex concepts, realistic characters, an intricate plot, an underlying history or circumstances, precise descriptions.

Table 4.8: Example task from our Amazon Mechanical Turk experiment.

HANNA V1 Summary. In total, including the human stories from the WRITINGPROMPTS dataset, HANNA V1 contains **1,056 stories** written or generated by **11 systems** (1 human writer, 3 ASG-specific systems, and 7 pretrained language models), aligned on **96 prompts**. Each story was annotated by three human raters and evaluated by the 72 automatic measures (including variants) selected in [Section 3.3](#). This corresponds to a total of **19,008 human annotations** and **76,032 automatic scores**.

Prompt: When you die, the afterlife is an arena where you face every insect and animal you killed in your life. If you win you go to heaven, lose you go to hell. Your job was an exterminator on earth.

Human: 3,000 years have I been fighting. Every morning, the raccoons scratch at my eyes. Every evening, the skunks spray me while the opossums chew at my feet. I have never had any tools. I have only my hands. I don't remember the place [...]

Story #1: First of all, not everyone was entitled to be an exterminator. But the ones that were – maybe were, like, *genius*, because, yes, I had once belonged to a less fortunate class of people – had all the opportunity to work for [...]

Story #2: It was hell. Not exactly a place of torture. There were no guards in prison and you couldn't just walk through it, either, because you would get killed regardless. hell was a young man, and he was lying on his floor. He was [...]

Story	RE	CH	EM	SU	EG	CX
Human	5	5	1	3	4	1
	2	2	3	2	2	3
	4	4	3	2	4	4
Story #1	2	4	3	1	1	1
	2	2	2	1	2	2
	2	3	2	3	3	3
Story #2	5	5	3	3	3	2
	3	2	3	2	2	3
	3	4	3	4	4	3

Metric	Human	Story #1	Story #2
BLEU ^{Ξ§} (%)	1.00	0.01	0.01
ROUGE-1 ^{Ξ§}	1.00	0.24	0.33
CHRF ^{Ξ§} (%)	1.00	0.32	0.39
BERTSCORE ^{Ξε}	1.00	0.50	0.52
MOVERSCORE ^{Ξε}	1.00	0.51	0.51
BARYSCORE ^{Ξε}	0.00	0.92	0.92
DEPTHSCORE	0.00	0.12	0.11
S3 ^{ΞΔ}	1.39	0.07	0.15
BARTSCORE ^{ΞΔ}	-0.98	-3.97	-4.03
SUPERT ^{Ξε}	0.94	0.37	0.36

Table 4.9: Example prompt, human and generated stories from HANNA with human annotations and measure scores.

4.3.2 Inter-Rater Reliability

To estimate the reliability of the annotations, we compute an inter-rater reliability (IRR) estimate for each criterion using the intra-class coefficient (ICC). Among the annotators who took part in the experiment, three of them covered 2,490 stories, *i.e.*, more than 78% of the

dataset. No annotator graded the same story twice. Since the reliability is to be estimated for the average of the three ratings, the ICC2k coefficient (ICC for *average random raters*) is the most relevant one, according to Hallgren [2012], as it incorporates the magnitude of the disagreement to compute its IRR estimate, with larger-magnitude disagreements resulting in lower ICC than smaller-magnitude disagreement. In particular, the ICC2k accounts for the systematic errors of raters and random residual errors.

We use the PINGOUIN³ library [Vallat, 2022] for computing the ICCs: results are shown in Table 4.10.

Coefficients are dispersed between 29% and 56% with relatively small confidence intervals (except for RE and CH), which can be considered between “fair” and “moderate” according to Landis and Koch [1977]. These values are in tune with existing literature [Graham et al., 2017; Habernal and Gurevych, 2017; Karpinska et al., 2021; Ritter et al., 2011; Spooren and Degand, 2010] and show the difficulty of evaluating natural language generation. Therefore, we follow the methodology of Artstein and Poesio [2008]; Craggs and Wood [2005], who argue against setting a specific agreement threshold as long as there is a detailed reporting of the methodology (see Table 4.8) and confidence intervals (Table 4.10).

Criterion	95% LB	ICC2k	95% UB
Relevance	0.18	0.48	0.65
Coherence	0.10	0.29	0.48
Empathy	0.25	0.34	0.41
Surprise	0.16	0.28	0.37
Engagement	0.34	0.46	0.55
Complexity	0.48	0.56	0.63

Table 4.10: Intra-class coefficient for average random raters per criterion. “95% LB” and “95% UB” stand for lower- and upper-bounds of the 95% confidence interval respectively. Higher is better.

4.3.3 Evaluating Our Human Criteria

Here, we study the relationships between our proposed human criteria. To compute overall (Figure 4.1) and system-level (Figure 4.2) absolute correlations, we average the human ratings. Following Mathur et al. [2020], who advise to remove outliers, we compute all correlations with human stories removed.

Overall Correlations (Figure 4.1). Kendall correlations range from 0.15 (RE–SU) to 0.62 (CH–EG), averaging at 0.40 globally. EG correlates slightly more with CH and CX; this corroborates the idea that coherent and intricate plots make readers more likely to connect with a story. In contrast, RE correlates on average poorly with the other criteria, which makes sense: an excellent story in every other aspect can be completely unrelated to a prompt, and vice versa.

³<https://pingouin-stats.org/>

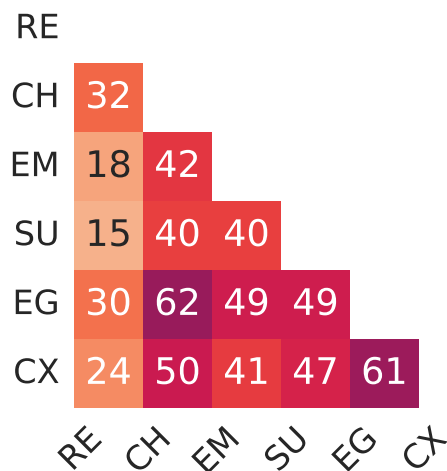


Figure 4.1: Overall absolute Kendall correlations between human criteria. Coefficient values are multiplied by 100 for readability; we will symbolize this with “($\times 100$)” in the next figures.

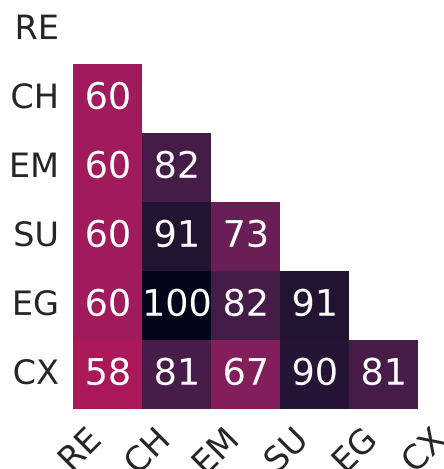


Figure 4.2: System-level absolute Kendall correlations ($\times 100$) between human criteria.

Overall, the moderate to weak correlations suggest that our criteria evaluate distinct aspects of storytelling which cannot be regrouped in fewer criteria.

System-level Correlations (Figure 4.2). Compared to overall correlations, system-level correlations are higher, ranging from 0.58 (RE–CX) to 1.00 (CH–EG). This suggests that a given system tends to be uniformly better or worse than other systems across all criteria. Apart from this observation, the criteria tend to behave similarly. In particular, RE has on average lower correlations with the other criteria.

Takeaways. The moderate to weak overall correlations between our human criteria are convincing evidence that they constitute a set of orthogonal criteria suited for human evaluation of ASG. At the system-level, we observe that current models are not specialized for a particular criterion: the high correlations suggest that a given system A tends to be uniformly better or worse than another system B across all criteria.

4.3.4 Comparing ASG Systems

Table 4.11 shows the average system ratings from our annotation experiment on HANNA V1 w.r.t. each of our human criteria. First, we observe that human stories achieve significantly higher scores than generated stories. Moreover, generic fine-tuned models perform better than ASG-specific systems, according to human annotators. The overall best system is GPT-2, which scores better than all other systems on all criteria. The GPT-2 variant trained with

Model	RE	CH	EM	SU	EG	CX	Average
Human	4.17 $\pm$ 0.14	4.43 $\pm$ 0.10	3.22 $\pm$ 0.14	3.15 $\pm$ 0.15	3.88 $\pm$ 0.12	3.73 $\pm$ 0.13	3.76 $\pm$ 0.06
BERTGEN	2.46 $\pm$ 0.16	3.14 $\pm$ 0.16	2.28 $\pm$ 0.13	2.09 $\pm$ 0.13	2.67 $\pm$ 0.12	2.41 $\pm$ 0.11	2.51 $\pm$ 0.06
CTRL	2.54 $\pm$ 0.16	2.93 $\pm$ 0.16	2.26 $\pm$ 0.13	1.93 $\pm$ 0.12	2.53 $\pm$ 0.12	2.23 $\pm$ 0.10	2.40 $\pm$ 0.06
GPT	2.40 $\pm$ 0.16	3.22 $\pm$ 0.15	2.37 $\pm$ 0.12	2.13 $\pm$ 0.13	2.76 $\pm$ 0.13	2.49 $\pm$ 0.12	2.56 $\pm$ 0.06
GPT-2	2.81 $\pm$ 0.16	3.29 $\pm$ 0.14	2.47 $\pm$ 0.12	2.21 $\pm$ 0.13	2.86 $\pm$ 0.12	2.68 $\pm$ 0.10	2.72 $\pm$ 0.06
GPT-2 (TAG)	2.67 $\pm$ 0.16	3.31 $\pm$ 0.15	2.47 $\pm$ 0.12	2.22 $\pm$ 0.13	2.92 $\pm$ 0.12	2.80 $\pm$ 0.11	2.73 $\pm$ 0.06
ROBERTA	2.54 $\pm$ 0.16	3.22 $\pm$ 0.16	2.27 $\pm$ 0.12	2.12 $\pm$ 0.13	2.74 $\pm$ 0.12	2.41 $\pm$ 0.11	2.55 $\pm$ 0.06
XLNET	2.39 $\pm$ 0.17	2.88 $\pm$ 0.16	2.10 $\pm$ 0.12	1.95 $\pm$ 0.12	2.46 $\pm$ 0.13	2.36 $\pm$ 0.11	2.36 $\pm$ 0.06
FUSION	2.09 $\pm$ 0.16	2.86 $\pm$ 0.16	1.99 $\pm$ 0.12	1.72 $\pm$ 0.12	2.27 $\pm$ 0.14	1.92 $\pm$ 0.11	2.14 $\pm$ 0.06
HINT	2.29 $\pm$ 0.16	2.38 $\pm$ 0.16	1.74 $\pm$ 0.13	1.56 $\pm$ 0.11	1.75 $\pm$ 0.12	1.45 $\pm$ 0.10	1.86 $\pm$ 0.06
TD-VAE	2.51 $\pm$ 0.16	2.99 $\pm$ 0.15	2.07 $\pm$ 0.11	2.10 $\pm$ 0.12	2.59 $\pm$ 0.12	2.49 $\pm$ 0.11	2.46 $\pm$ 0.06

Table 4.11: Average system ratings per criterion with 95% confidence interval. Best value in **bold blue**, values in the confidence interval of the best value in **bold** without asterisk. Higher is better.

<EOP> tags shows marginal improvement compared to GPT-2, as reported in Bai et al. [2021]. It is worth noting that all models are still noticeably below human performance.

4.3.5 Need for a HANNA Upgrade

Soon after we released HANNA V1, the Natural Language Processing field underwent yet another series of advances spearheaded by the rise of Large Language Models (LLM) such as GPT-3 [Brown et al., 2020], LAMDA [Thoppilan et al., 2022], PALM [Chowdhery et al., 2023] and LLAMA [Touvron et al., 2023a]. These new models have been setting new state-of-the-art performance standards for a wide array of NLP tasks, e.g. question answering, summarization, and machine translation. In particular, for ASG, LLMs are now able to produce convincing stories, so much so that they can be hard to distinguish from human stories [Clark et al., 2021]. As their performance improves, they may become valuable assistants to our creative process; already, writing contests have been shown to encourage their use [Edilivre, 2023].

4.4 Augmenting HANNA with Large Language Models

The increased availability of LLMs to the general public underlines the need for reliable story evaluation methods that can be used to improve both the performance of ASG models and our understanding of their strengths and weaknesses. Since the human annotation of stories is costly and time-consuming [Celikyilmaz et al., 2020], Automatic Story Evaluation (ASE) using LLMs could provide an efficient and scalable replacement for human evaluation.

In this section, we use LLMs to produce new stories for HANNA and perform several annotation experiments where we ask LLMs to rate stories according to different human criteria and to explain their rating. We show an example of Eval-prompt and answer in Figure 4.3.

Eval-Prompt

Prompt: A mirror shows your reflection and your future soulmate. You only see your reflection.

Story: Carnival music seemed to dance in the background. A cacophony of color and knick knacks decorated the fairgrounds. It was just another day at the carnival. I sit at my stand reading another old junky sci-fi novel. "So how much to see it?" Some pre-teen girl interrupted my book with her piercing voice. I sigh towards her and lean in. "Ten dollars, just like the sign says." "And it works?" "So I'm told." "You mean you've never looked?" "Nope," I say, picking my book back up. "Well why not." "Because some things in this life should be a mystery," I reply. "Well I want to see," she says. I tap on the money box and she places her ten dollars in. I watch her go in and wait for the audible reaction. After a couple of seconds I can hear an excited squeal. She runs out jumping up and down. "He's so handsome!" I can see her run off in excitement. I smirk. Another idiot, I think to myself. The carnival director walks up shortly after. "Time to start packing up, Isaac," he says. I nod and begin to break down my stand. My table is put away, and the money is locked up now. It's that time again to put away the mirror. I stand in front of it like every other day. It's just me, again. Same as yesterday, and the day before that, I am alone in the reflection. I smile at my reflection as I cover the mirror. I put the tent and everything in the back of my trailer with a skip in my step. You have to feel sorry for these suckers, I think. They're destined for one person only. Who knows if they'll ever find them? At least I know where to always find myself.

Rate the story on a scale from 1 to 5 on Empathy and explain your answer.

Rating:

LLM Answer

I would rate the story a 2 on Empathy. While I could relate to the character's feelings of indifference towards the magic mirror, I did not feel a strong emotional connection to any of the characters. There were no specific emotions that I recognized in the characters, and I could not completely relate to them on an emotional level.

Figure 4.3: Example Eval-Prompt and answer from our experiments. "Prompt" inside the Eval-Prompt refers to the story-prompt.

We release this second version of HANNA, which we will call HANNA V2, on GitHub⁴.

⁴<https://github.com/dig-team/hanna-benchmark-asg>

Our data consists of:

- **ASG experiment:** 480 stories generated by Llama models with corresponding LLM annotations to expand the HANNA corpus;
- **ASE experiments:** ~150k rating and explanation annotations (1,056 stories, 6 criteria, 4 Eval-Prompts, 3 tries, 2 models).

4.4.1 Experimental Protocol for ASG using LLMs

Prompt		
Write a story about an elderly wizard and his young female apprentice.		
Llama-7B Story	Beluga-13B Story	Platypus2-70B Story
It was a crisp autumn day when the elderly wizard, Orion, welcomed his new apprentice, Lily, to his tower. Orion had been searching for a worthy successor for many years, and he had finally found the perfect person in Lily. She was a bright and curious young woman, eager to learn the ways of magic. As Lily entered the tower, she was struck by its grandeur. It was filled with ancient tomes, dusty artifacts, and mysterious creatures. [...]	In the enchanted realm of Cascadia, a wise old wizard named Elder Oak lived in solitude, guarding the ancient forest's secrets and magic. The trees seemed to whisper, and the winds sighed their stories to him, for he was the last of his kind. He spent his days learning and protecting the natural wisdom, and his nights were filled with dreams of a future where the wonders of his world were cherished and honored. [...]	Once upon a time, in the far-away magical kingdom of Alvoraria, there lived an elderly wizard named Azmandor. Azmandor had dedicated his entire life to the pursuit of knowledge, particularly in the field of magic. His quest had brought him great fame and recognition among his peers, but little happiness. Now, as he approached the twilight of his life, he longed for a new apprentice to pass on his vast store of magical knowledge. [...]

Table 4.12: Example aligned stories from Llama-7B, Beluga-13B, and Platypus2-70B.

Since the release of the HANNA V1 dataset, language models have made significant advancements. We therefore feel the need to expand HANNA with stories generated by more recent models and, namely, Large Language Models (LLMs). We consider the source-available Llama2 models [Touvron et al., 2023b] for that purpose.

We select Llama-2-7b-chat-hf (Llama-7B) as a new baseline and 4 high-performing models (at the time of selection) of different sizes on the HuggingFace Open LLM Leaderboard⁵:

1. Platypus2-70B-instruct (Platypus2-70B),
2. llama-30b-instruct-2048 (Llama-30B),
3. StableBeluga-13B (Beluga-13B),
4. Mistral-7B-OpenOrca (Mistral-7B).

We use the following LLM-prompt, where “PROMPT” is replaced by a given story-prompt:

⁵https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Human: Write a story based off the prompt given below.

Prompt: {PROMPT}

Assistant:

As a result, we obtain 480 new stories generated with 5 Large Language Models. See Table 4.12 for an example of aligned stories. We can observe that, compared with older ASG-specific models (Table 4.7), the stories generated by LLMs appear to be much more fluent.

4.4.2 Experimental Protocol for Automatic Annotation using LLMs

In this section, we use LLMs as Automatic Story Evaluation measures. We will refer to the prompt that is fed to the LLM as the *Eval-Prompt*, to distinguish it from the story-prompt. See Figure 4.3 for an example of the use of an LLM for story evaluation and Figure 4.4 for a schematic overview of our experiments.

■ Related work

Automatic Annotation with LLMs. LLMs are increasingly being tested for automatic text annotation, *e.g.* for sentiment analysis [Qureshi et al., 2022], named entity recognition [Enkh-saikhan et al., 2021] or event structure modeling [Vauth et al., 2021]. Wang et al. [2021] propose to leverage GPT-3 as a data labeler and show that it can be up to 96% more cost-efficient than human labeling on a variety of NLP tasks. They also explore various strategies of mixing labeled data from GPT-3 and humans under a fixed budget and achieve better performance than using data from a single labeler. Ding et al. [2023] conduct a comprehensive analysis on using GPT-3 for data annotation for complex NLP tasks. They find that directly annotating unlabeled data is suitable for tasks with small label space (*e.g.* binary stance classification) while generation-based methods are more suitable for tasks with large label space (*e.g.* neural entity recognition). Overall, they find that generation-based approaches tend to be more cost-effective compared with human annotation of unlabeled data, but that GPT-3 annotations are still of inferior quality compared with human annotations. Chakrabarty et al. [2024] design the Torrance Test for Creative Writing (TTCW), a set of 14 tests based on the four original Torrance dimensions of fluency, flexibility, originality, and elaboration. They experimentally validate the TTCW through an assessment of 48 stories written by humans or generated by LLMs and involving 10 participants with expertise in creative writing. They find that the experts reach moderate agreement on individual dimensions, and strong agreement when evaluating all dimensions in aggregate. They show that LLM-generated stories are three to ten times less likely to pass TTCW tests compared to expert-written stories, and that current state-of-the-art LLMs are not yet capable of reproducing expert assessments when administering TTCW tests. We seek to generalize their findings through a larger-scale evaluation (1,000+ stories), a more diverse pool of LLMs, including source-available models, and a finer analysis and discussion of LLM performance.

Prompt Engineering. The importance of designing efficient prompts for large language models such as GPT-3 has been extensively investigated in recent years. A systematic survey of prompt engineering techniques can be found in [Sahoo et al. \[2024\]](#): they explain that zero-shot and few-shot prompting were the first paradigms that emerged with LLMs. Those setups were generally sufficient, except for more reasoning-oriented tasks, to which LLMs failed more often. Chain-of-Thought prompting [[Wei et al., 2022b](#)] was therefore developed, soon followed by multiple variants. [Radford et al. \[2019\]](#) use zero-shot prompting for benchmarking their GPT-2 model. This technique removes the need for extensive training data, instead relying on carefully crafted prompts that guide the model toward novel tasks. They find that GPT-2 zero-shots to state-of-the-art performance on 7 out of 8 tested language modeling datasets, and they surmise that high-capacity models trained to maximize the likelihood of a sufficiently varied text corpus become unsupervised multitask learners. [Brown et al. \[2020\]](#) introduce the GPT-3 language model and test its performance in a few-shot setting, where the model is given a few input-output examples to help it grasp the task at hand. They demonstrate that, in contrast to zero-shot prompting, a few high-quality examples are enough to enhance model performance on complex tasks. However, few-shot prompting could be less adapted to longer text inputs, as it needs extra tokens for the examples. Furthermore, they discover that biases like favoring frequent terms may have an impact on few-shot results, and that the design of the provided examples greatly influence model behavior. [Reynolds and McDonell \[2021\]](#) notably find that zero-shot prompting can perform similarly to few-shot prompting, and even exceed it. They explore the design of metaprompts that prime the language model to better solve a given problem. [Zhou et al. \[2023b\]](#) treat the prompt engineering process as an optimization problem, use search algorithms guided by LLMs to solve it and attain human-level performance. [Wei et al. \[2022b\]](#) introduce Chain-of-Thought (CoT), a prompting strategy that enables models to decompose complex problems into simpler steps. Through a series of experiments, they demonstrate that CoT prompting improve model performance for both commonsense and symbolic reasoning. They achieve state-of-the-art results in math and commonsense reasoning benchmarks with PALM 540B, reaching an accuracy of 90.2%. [Wei et al. \[2022a\]](#) and [White et al. \[2023\]](#) review different strategies that have been applied to augment large language model abilities, *e.g.* least-to-most prompting [[Zhou et al., 2023a](#)], ask-me-anything prompting [[Arora et al., 2023](#)], and zero-shot chain-of-thought reasoning [[Kojima et al., 2022](#)].

In our setting, we choose to investigate whether LLMs perform better with simple or detailed guidelines, and with zero- or one-shot Eval-Prompts.

■ ASE Criteria

We re-use our set of six human criteria introduced in [Section 3.2.3](#):

1. **Relevance** (RE, how well the story matches its prompt),
2. **Coherence** (CH, how much the story makes sense),
3. **Empathy** (EM, how well the reader understood the character’s emotions),
4. **Surprise** (SU, how surprising the end of the story was),

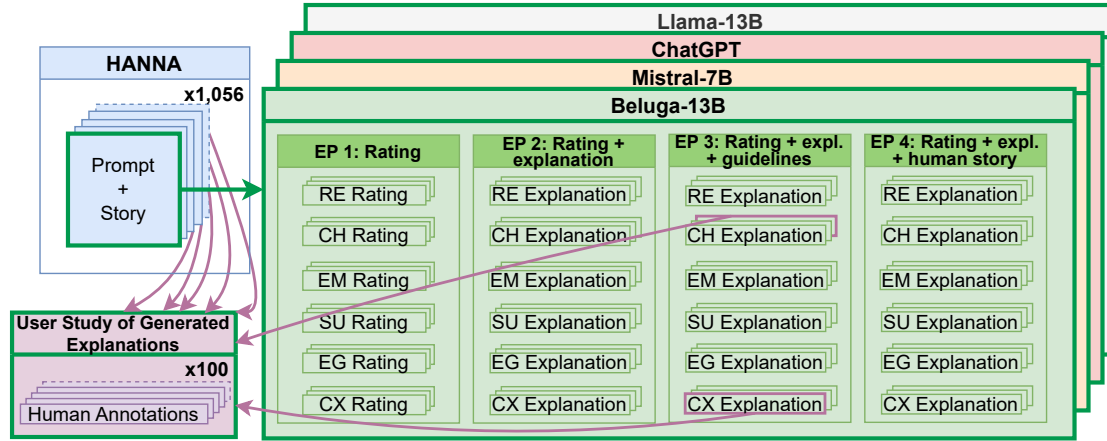


Figure 4.4: Schema of the performed ASE experiments. RE, CH, etc. are the considered human criteria. “EP” means “Eval-Prompt”, defined in [Section 4.4.2](#). For the user study ([Section 6.3.3](#)), we randomly sampled 100 explanations from our experiments.

5. **Engagement** (EG, how much the reader engaged with the story),
6. **Complexity** (CX, how elaborate the story is).

■ Methodology

Given the importance of good prompt engineering [[Zhao et al., 2021](#)], we design four different Eval-Prompts for the generation of ratings. For each of our Eval-Prompts, we provide the model with a story-prompt and a corresponding story. Then:

1. **Eval-Prompt 1** (simple rating): we ask the model to rate the story on a scale from 1 to 5 on one of the six criteria;
2. **Eval-Prompt 2** (rating with explanation): same as Eval-Prompt 1, and we ask the model to explain its answer;
3. **Eval-Prompt 3** (rating with explanation and guidelines): same as Eval-Prompt 2, and we provide the model with the detailed guidelines from our original human annotation protocol (see [Section 4.3.1](#));
4. **Eval-Prompt 4** (rating with explanation and human story): same as Eval-Prompt 2, and we provide the model with the human story associated with the same story-prompt. We explicitly tell the model that the human story is only given for reference purposes.

Different Eval-Prompt examples can be found in [Figures 4.5, 4.6, 4.7, and 4.8](#).

Eval-Prompt 1

Prompt: You have become death, destroyer of worlds.

Story: You look up to see all of them in fear. You just must fix this soon. Slowly, just like your Father always had instructed him, you look down and see all your foes dead and beaten down. You can't resist the urge to touch the wounds. For there is nothing you can do about it. [...]

Rate the story on a scale from 1 to 5 on Surprise (how surprising the end of the story was).

Rating:

Figure 4.5: Example of an Eval-Prompt 1 for the Surprise criterion.

Eval-Prompt 2

Prompt: You have become death, destroyer of worlds.

Story: You look up to see all of them in fear. You just must fix this soon. Slowly, just like your Father always had instructed him, you look down and see all your foes dead and beaten down. You can't resist the urge to touch the wounds. For there is nothing you can do about it. [...]

Rate the story on a scale from 1 to 5 on Surprise (how surprising the end of the story was) and explain your answer.

Rating:

Figure 4.6: Example of an Eval-Prompt 2 for the Surprise criterion.

Eval-Prompt 3

Prompt: You have become death, destroyer of worlds.

Story: You look up to see all of them in fear. You just must fix this soon. Slowly, just like your Father always had instructed him, you look down and see all your foes dead and beaten down. You can't resist the urge to touch the wounds. For there is nothing you can do about it. [...]

Guidelines:

- 1 — The ending seemed completely obvious from the start, or doesn't make any sense at all.
- 2 — The ending was easily predictable after a few sentences.
- 3 — The ending was predictable after half of the story.
- 4 — The ending surprised you, but would have been difficult to predict.
- 5 — The ending surprised you, and still seemed as if it could very reasonably have been predicted, ie, there were enough clues in the story.

Rate the story on a scale from 1 to 5 on Surprise (how surprising the end of the story was) and explain your answer. Use the provided guidelines.

Rating:

Figure 4.7: Example of an Eval-Prompt 3 for the Surprise criterion.

Eval-Prompt 4

Prompt: You have become death, destroyer of worlds.

Target Story: You look up to see all of them in fear. You just must fix this soon. Slowly, just like your Father always had instructed him, you look down and see all your foes dead and beaten down. You can't resist the urge to touch the wounds. For there is nothing you can do about it. [...]

Human Story: I saw the button. It was simple, red, no words on it as I already knew what it did. I mean I built the button, I built what happens when you press the button, and I was given the choice of whether or not to push the button. Humanity was screwed, that much was apparent, Plagues and starvation across 3 planets and 6 space stations. [...]

Rate the target story on a scale from 1 to 5 on Surprise (how surprising the end of the story was) and explain your answer. Do not rate the human story; it is here only for reference.

Rating:

Figure 4.8: Example of an Eval-Prompt 4 for the Surprise criterion.

■ Automatic Story Evaluation Models

We consider source-available Llama-2 [Touvron et al., 2023b] models for our experiments, and the closed-source gpt-3.5-turbo model from OpenAI. Precisely, we use the 5 following models:

1. StableBeluga-13B (Beluga-13B),
2. Llama-2-13b-chat-hf (Llama-13B),
3. Mistral-7B-OpenOrca (Mistral-7B),
4. Llama-2-7b-chat-hf (Llama-7B),
5. gpt-3.5-turbo (ChatGPT).

Llama-2 models were trained on a closed “new mix of data from publicly available sources”. Beluga-13B and Mistral-7B are Llama-2 models fine-tuned on Orca-style datasets which contain triplets of “System message–User query–LLM response” for a large collection of tasks [Mukherjee et al., 2023]. Beluga-13B is fine-tuned on StabilityAI’s closed internal dataset, while Mistral-7B is fine-tuned on the open OpenOrca dataset [Lian et al., 2023].

ChatGPT [Brown et al., 2020; Ouyang et al., 2022] is a closed-source model trained on a closed internal dataset that includes the CommonCrawl, Books1 and Books2 datasets.

We submit each of the four Eval-Prompts 3 times on all 1,056 HANNA stories on each of the 6 criteria, and we then extract the ratings automatically from the generated answer via a regular expression. Since story evaluation on multiple prompts and multiple criteria is computationally demanding, we limit our experiments to 13B and 7B models.

Llama-7B failed at the task too often for the results to be exploitable, *e.g.* by generating nonsensical conversations between itself and the user: in more than half of its answers, no rating was provided at all, so we excluded it from our analysis.

We use the default parameter values, *i.e.*, (temperature, top_p) = (1, 0.95) for Llama models and (0.7, 1) for ChatGPT.

■ Libraries

We use the TRANSFORMERS library [Wolf et al., 2020] and the OpenAI API for our experiments.

Model	RE	CH	EM	SU	EG	CX	Average
Human	3.37 ±0.12	3.55 ±0.11	3.42 ±0.11	3.11 ±0.13	3.58 ±0.10	3.48 ±0.10	3.42 ±0.06
Platypus2-70B	4.09 ±0.05	4.31 ±0.05	3.92 ±0.06	3.69 ±0.07	4.19 ±0.05	3.88 ±0.05	4.01 ±0.03
Llama-30B	4.19 ±0.05	4.38 ±0.04	4.04 ±0.06	3.63 ±0.09	4.31 ±0.05	3.98 ±0.05	4.08 ±0.03
Beluga-13B	4.06 ±0.08	4.10 ±0.06	3.75 ±0.08	3.54 ±0.08	3.90 ±0.08	3.69 ±0.07	3.84 ±0.05
Mistral-7B	4.12 ±0.05	4.25 ±0.05	3.86 ±0.06	3.56 ±0.08	4.11 ±0.05	3.82 ±0.04	3.95 ±0.03
Llama-7B	4.07 ±0.06	4.24 ±0.05	3.90 ±0.06	3.58 ±0.06	4.09 ±0.05	3.79 ±0.05	3.95 ±0.03
GPT-2	2.57 ±0.13	2.36 ±0.11	2.72 ±0.11	2.59 ±0.14	2.67 ±0.12	2.89 ±0.12	2.63 ±0.07
HINT	1.57 ±0.10	1.31 ±0.07	1.59 ±0.10	1.49 ±0.10	1.58 ±0.09	1.43 ±0.08	1.49 ±0.06

Table 4.13: Average Beluga-13B ratings for Eval-Prompt 1 with 95% confidence interval. Higher is better.

Model	RE	CH	EM	SU	EG	CX	Average
Human	3.48 $\pm$ 0.11	3.50 $\pm$ 0.10	3.69 $\pm$ 0.08	3.24 $\pm$ 0.11	3.42 $\pm$ 0.10	3.45 $\pm$ 0.07	3.46 $\pm$ 0.05
Platypus2-70B	4.26 $\pm$ 0.08	4.31 $\pm$ 0.08	4.05 $\pm$ 0.07	3.46 $\pm$ 0.10	3.94 $\pm$ 0.06	3.55 $\pm$ 0.07	3.93 $\pm$ 0.03
Llama-30B	4.15 $\pm$ 0.10	4.29 $\pm$ 0.07	4.02 $\pm$ 0.07	3.46 $\pm$ 0.09	3.94 $\pm$ 0.06	3.65 $\pm$ 0.07	3.92 $\pm$ 0.03
Beluga-13B	4.07 $\pm$ 0.09	4.14 $\pm$ 0.07	3.98 $\pm$ 0.07	3.50 $\pm$ 0.09	3.74 $\pm$ 0.08	3.59 $\pm$ 0.07	3.84 $\pm$ 0.03
Mistral-7B	4.15 $\pm$ 0.10	4.22 $\pm$ 0.08	4.02 $\pm$ 0.07	3.51 $\pm$ 0.11	3.94 $\pm$ 0.07	3.67 $\pm$ 0.07	3.92 $\pm$ 0.04
Llama-7B	4.13 $\pm$ 0.10	4.14 $\pm$ 0.09	3.90 $\pm$ 0.08	3.48 $\pm$ 0.09	3.78 $\pm$ 0.08	3.56 $\pm$ 0.08	3.83 $\pm$ 0.05
GPT-2	2.40 $\pm$ 0.10	2.37 $\pm$ 0.09	2.74 $\pm$ 0.10	2.85 $\pm$ 0.11	2.60 $\pm$ 0.09	2.88 $\pm$ 0.09	2.64 $\pm$ 0.05
HINT	2.12 $\pm$ 0.11	2.13 $\pm$ 0.08	2.23 $\pm$ 0.10	2.28 $\pm$ 0.11	2.05 $\pm$ 0.08	2.05 $\pm$ 0.09	2.15 $\pm$ 0.06

Table 4.14: Average Mistral-7B ratings for Eval-Prompt 1 with 95% confidence interval. Higher is better.

4.4.3 LLM Performance at ASG

In Table 4.13 and Table 4.14, we display the average ratings attributed by respectively Beluga-13B and Mistral-7B to multiple V1 and V2 systems. For the sake of readability, we restrict the scope of the tables to human stories, LLM stories and GPT-2 and HINT, respectively the best and worst systems from HANNA V1. Other V1 systems have similar ratings to the latter two.

We observe that, according to Beluga-13B and Mistral-7B, LLMs perform remarkably well, getting higher ratings than older models (GPT-2) and even human stories. Beluga-13B and Mistral-7B both seem to prefer the outputs from larger LLMs (Platypus2-70B, Llama-30B) to their own outputs, suggesting that the LLM grading process cannot be explained simply by a proxy for perplexity. Interestingly, in both tables, Mistral-7B gets slightly higher ratings than Beluga, with some differences being statistically discernible, which could be explained by differences in fine-tuning data.

Takeaways. Larger models (Platypus2-70B, Llama-30B) exhibit the best ASG performance, with LLM ratings at least equal to those of human stories. However, our setting involves short stories of between 500 and 1,000 words; generating longer stories may prove more difficult since maintaining large-scale coherence may become an issue. Moreover, we need to confirm that Automatic Story Evaluation performed by LLMs is reliable proxy for human evaluation: we tackle this question in Section 5.2.

4.5 Conclusion

Motivated by our survey of the Automatic Story Generation literature (Section 4.2), we built HANNA, a new corpus designed for ASG evaluation. The first version of HANNA contains 1,056 stories written or generated by 11 systems (1 human writer, 3 ASG-specific systems, and 7 pretrained language models), and aligned on 96 prompts (Section 4.3.1) We performed an annotation experiment on HANNA V1, asking human raters to grade each story w.r.t. our six human criteria. We computed intra-class coefficients to estimate the consistency of our 19,0008 annotations and found that our values are in tune with the existing NLP literature (Section 4.3.2).

Furthermore, we analyzed the overall and system-level correlations between our human criteria, and confirmed that they allow for a standardized and extensive human evaluation (Section 4.3.3). Indeed, our six criteria are moderately to weakly correlated with one another, which shows that they are non-redundant, and produce coherent system rankings. Additionally, we observed that large pre-trained language models produce the best results for ASG (Section 4.3.4). In particular, GPT-2 performed better than systems specifically tailored for ASG despite being older than some of them. Overall, all systems remained significantly inferior to human output.

Finally, we used Large Language Models to augment HANNA with 480 new stories (Section 4.4.1) and 150k+ annotations containing ratings of stories w.r.t. different Eval-Prompts and human criteria (Section 4.4.2). We also observed that LLMs obtain better grades than humans when evaluated by two different LLMs: Beluga-13B and Mistral-7B (Section 4.4.3).

In Chapter 5, we conduct our meta-evaluation benchmark of automatic measures on HANNA and evaluate the performance of Large Language Models on Automatic Story Evaluation.

Chapter 5

Meta-Evaluation Benchmark of Automatic Story Generation

*“ Whether or not you enjoy
something simply depends on your
own heart. ”*

Machine Receptionist
(NieR:Automata)

Contents

5.1	Meta-Evaluation on HANNA V1	96
5.1.1	Evaluation of Automatic Measures	96
5.1.2	Fine-Grained Analysis	102
5.1.3	Takeaways	110
5.2	Large Language Models for Automatic Story Evaluation	111
5.2.1	ASE1: Comparison with Current Evaluation Measures	111
5.2.2	ASE2: Influence of the Eval-Prompt	116
5.3	Conclusion	118
5.3.1	Takeaways	118
5.3.2	Limitations and Future Directions	118

Chapter 5 Abstract

After we have described the methodology for our meta-evaluation of ASG (Chapter 3) and built a corpus specifically designed for that task (Chapter 4), this chapter details our meta-evaluation of automatic evaluation measures for Automatic Story Generation. First, we present our meta-evaluation of non-LLM automatic measures (Section 5.1.1) on HANNA V1, and a more fine-grained analysis of measures using specific methods (Section 5.1.2). We mainly observe that specific measures for Automatic Story Evaluation are needed, and that commonly used measures such as BLEU are sub-optimal. We then show our analysis of LLM performance at ASE: we compare LLMs to human and non-LLM automatic evaluation methods (Section 5.2.1) and we examine the influence of the Eval-Prompt on LLM performance (Section 5.2.2). We find that LLMs are currently the best proxy for human evaluation of ASG and that, in our specific setting, providing detailed guidelines does not improve correlations between LLM and human ratings.

We will begin with the experimental results we obtained on HANNA V1 (Section 5.1), and then proceed with our experiments using Large Language Models (Section 5.2).

5.1 Meta-Evaluation on HANNA V1

5.1.1 Evaluation of Automatic Measures

In this section, we evaluate how suitable existing automatic measures are for Automatic Story Evaluation. We use the SUMMEVAL library [Fabbri et al., 2021]¹. For the sake of readability, in each figure, we selected representative measures from each of the categories introduced in Section 3.3.1. Full figures can be found in Section A.2.

■ Correlations With Human Judgement

Overall Correlations (Figure 5.1). Most automatic evaluation measures have either a moderate (between 0.30 and 0.50) or weak (below 0.30) correlation with human criteria. RE is particularly elusive, except for the SUPERT measure, which is reference-free and compares the prompt and the story. This corroborates Novikova et al. [2017], who argue that automatic measures do not accurately reflect human judgment when comparing instances despite performing reliably at the system-level. We also observe vertical “color stripes”: measure performance seems to be consistent across criteria. A weak measure will correlate poorly with all criteria, whereas a more robust measure will be uniformly better. Automatic measures tend to correlate more highly on average with CX, which could be explained by the tendency of annotators to liken story complexity to syntactic complexity.

¹<https://github.com/Yale-LILY/SummEval>

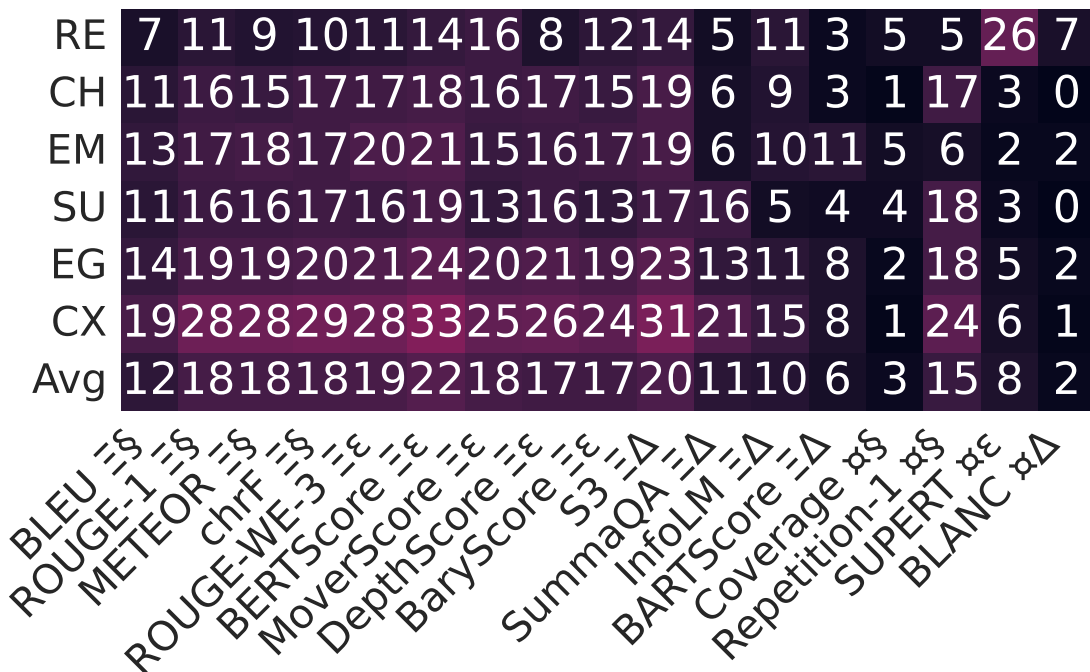


Figure 5.1: Overall absolute Kendall correlations ($\times 100$) between automatic measures and criteria. Full version shown on [Figure A.1](#).

System-level Correlations (Figure 5.2). Similarly to human criteria ([Section 4.3.3](#)), correlations are higher at the system-level, hovering between 0.40 and 0.70 for most measures. Therefore, while automatic measures are poor estimators of human criteria for a specific story, they can be used to compare systems with reasonable accuracy. In addition, contrary to expectations that story evaluation should not require a reference, reference-free measures (e.g. SUPERT, BLANC) exhibit noticeably lower correlation values.

Best Measures per Criterion (Table 5.1 and Table 5.2). For overall correlations, data statistics such as TEXT LENGTH and REPETITION outperform the majority of automatic measures for Surprise, Engagement and Complexity, which highlights the shortcomings of current automatic measures. For Relevance, SUPERT performs best, which may be explained by the fact that it directly compares the prompt and the story. Finally, among reference-based measures, S3 and BERTSCORE are among the best performers.

At the system-level, BARYSCORE is the best performing measure for all criteria except Relevance. Changing the hyperparameters only weakly affect the correlation values, which suggests that its evaluation strategy based on Wasserstein distributions captures system-level features which are closer to human criteria than other automatic measures.



Figure 5.2: System-level absolute Kendall correlations ($\times 100$) between automatic measures and criteria. Full version shown on [Figure A.2](#).

■ Correlations Between Automatic Measures

Overall correlations (Figure 5.3). Reference-based measures are poorly to highly correlated with one another, with values ranging from 0.07 to 0.80. Interestingly, text-based measures correlate better w.r.t. one another, and so do embedding-based-measures. This effect is less apparent for model-based measures.

By contrast, reference-free measures, and especially embedding- and model-based reference-free measures such as SUPERT and BLANC correlate very poorly w.r.t. all other measures, even other reference-free measures. This emphasizes the fact that reference-free measures can use widely different and unrelated evaluation strategies.

System-level Correlations (Figure 5.4). Previous observations for overall correlations remain mostly valid, although correlations are overall higher. Reference-based measures form a large group of very highly correlated measures, with a majority of correlations surpassing 0.70. Reference-free measures remain more weakly correlated to other measures: most correlations are less than 0.30.

Crit.	#1 Measure	τ	#2 Measure	τ	#3 Measure	τ
Relevance	SUPERT-SS	0.26	SUPERT-PS	0.26	SUPERT-GOLD	0.17
Coherence	REPETITION-3	0.20	S3 _R	0.19	S3 _P	0.19
Empathy	BERTSCORE _R	0.21	ROUGE-WE-3 _R	0.20	S3 _P	0.19
Surprise	TEXT LENGTH	0.20	REPETITION-3	0.19	BERTSCORE _R	0.19
Engagement	BERTSCORE _R	0.24	TEXT LENGTH	0.23	S3 _P	0.23
Complexity	TEXT LENGTH	0.35	BERTSCORE _R	0.33	S3 _P	0.31

Table 5.1: Top 3 measures per criterion for overall Kendall (τ) correlation values. Indices denote different variants.

Crit.	#1 Measure	τ	#2 Measure	τ	#3 Measure	τ
Relevance	S3 _P	0.60	CHRF	0.60	ROUGE-SU* _R	0.60
Coherence	BARYSCORE _{0.001}	0.78	BARYSCORE ₅	0.69	BARYSCORE ₁₀	0.69
Empathy	BARYSCORE _{0.01}	0.78	BERTSCORE _F	0.73	BARYSCORE _{0.01}	0.73
Surprise	BARYSCORE _{0.001}	0.78	BARYSCORE ₅	0.69	BARYSCORE ₁₀	0.69
Engagement	BARYSCORE _{0.001}	0.78	BARYSCORE ₅	0.69	BARYSCORE ₁₀	0.69
Complexity	BARYSCORE ₅	0.76	BARYSCORE ₁₀	0.76	NOVELTY-1	0.72

Table 5.2: Top 3 measures per criterion for system-level Kendall (τ) correlation values. Indices denote different variants.

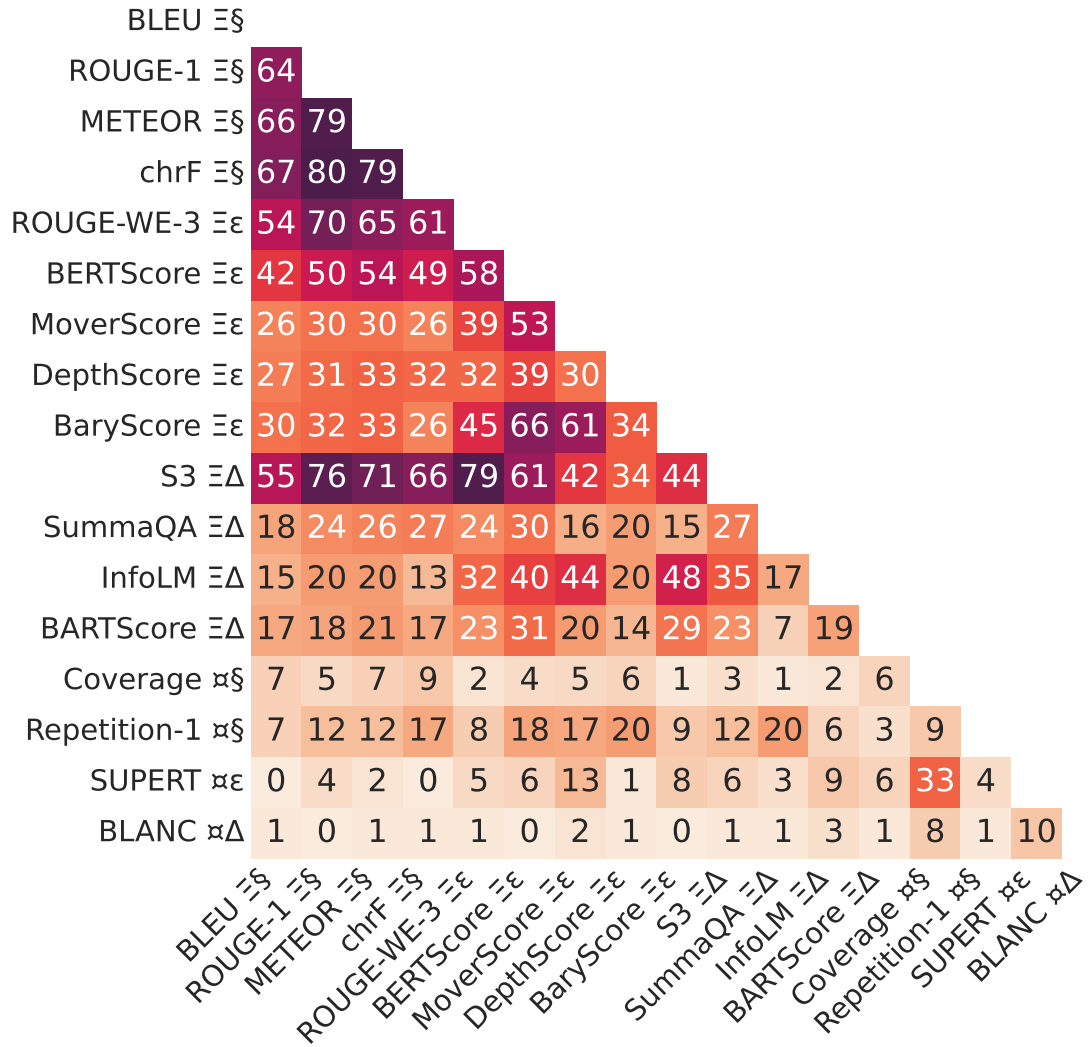


Figure 5.3: Overall absolute Kendall correlations ($\times 100$) between automatic measures. Full version shown on [Figure A.3](#).

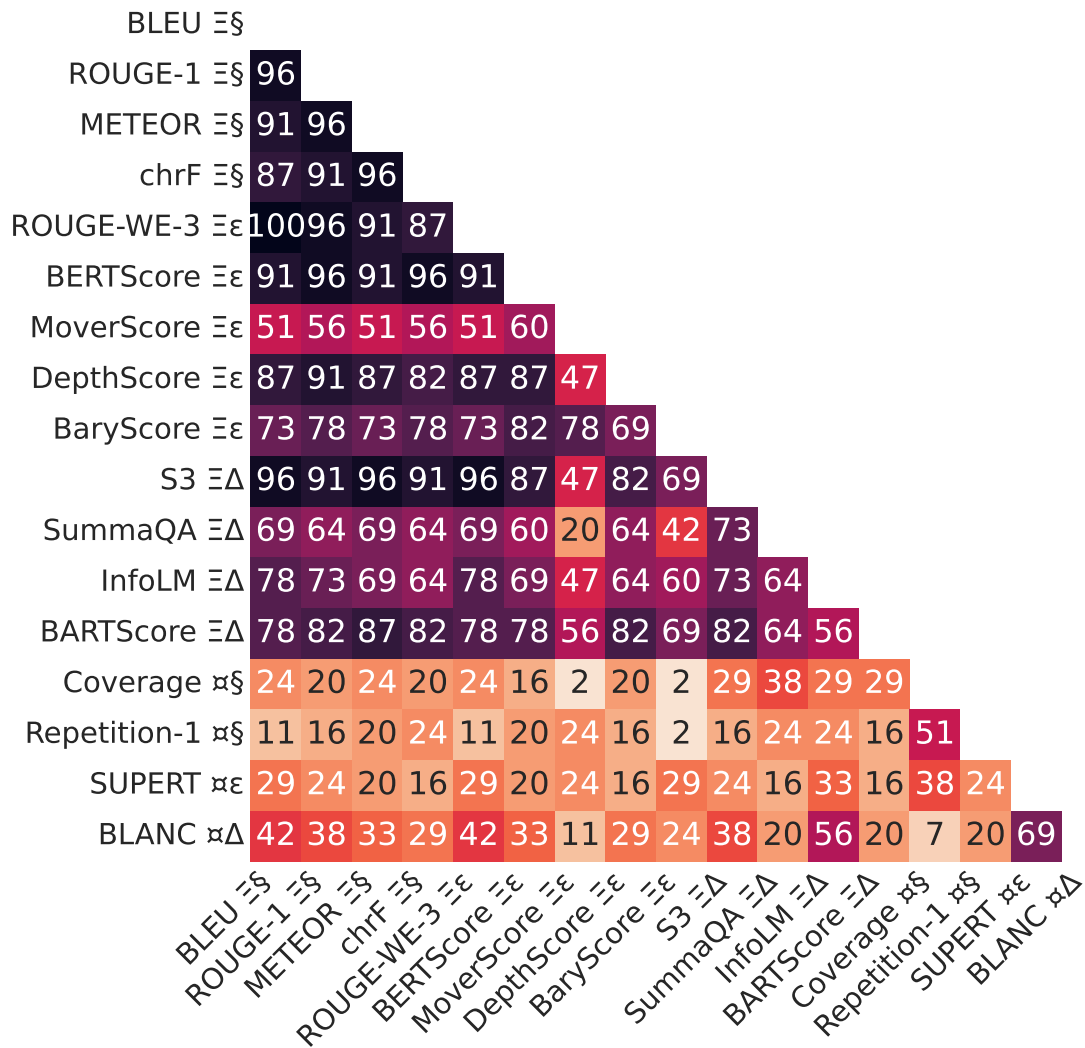


Figure 5.4: System-level absolute Kendall correlations ($\times 100$) between automatic measures. Full version shown on [Figure A.4](#).

5.1.2 Fine-Grained Analysis

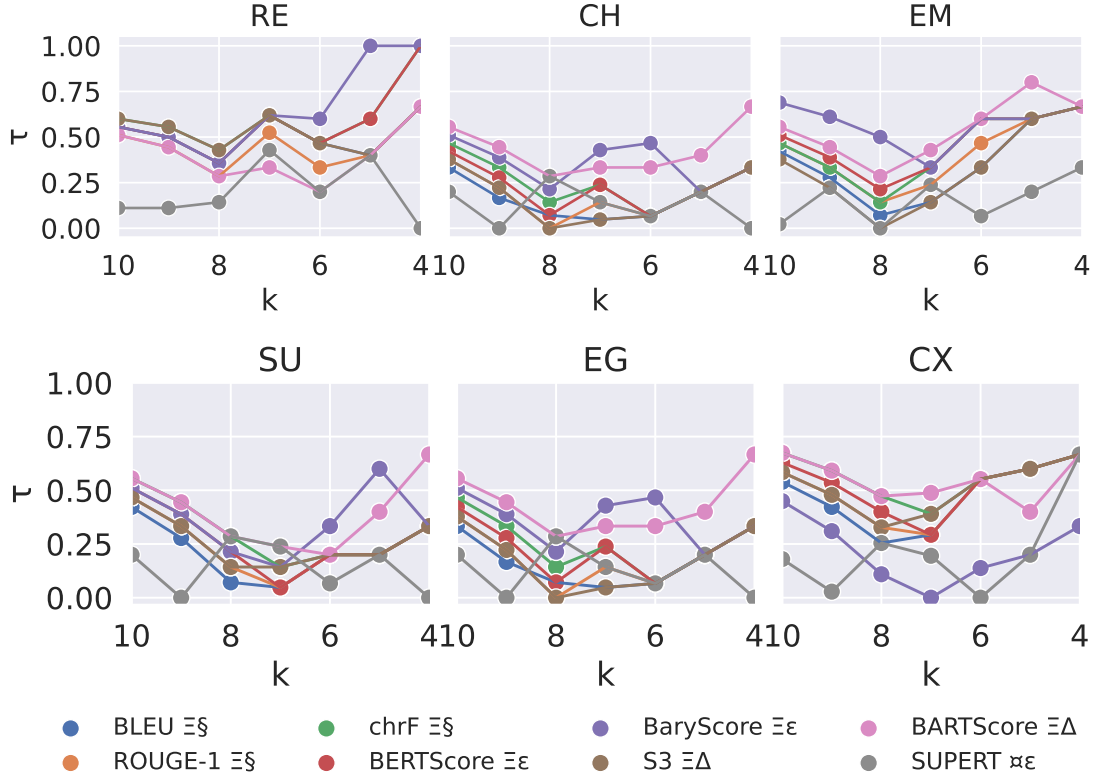


Figure 5.5: System-level absolute Kendall correlations between automatic measures and our proposed human criteria on top- k systems.

■ Top- k Systems (Figure 5.5)

Here, we evaluate whether automatic measures can reliably quantify differences between systems depending on their relative performance w.r.t. one another. For all criteria, except RE, correlations tend to follow a convex curve between $k = 10$ and $k = 4$, suggesting that measures should not be used to compare systems of high variance in quality unless there are enough of them. Indeed, removing a few systems causes correlations to worsen significantly, until the remaining systems are few enough and of competitive performance. RE correlations interestingly increase as k decreases, which indicates that system quantity is a lesser concern for RE.

■ Pairwise System Comparison (Figure A.5)

Here, we evaluate the pairwise discriminative power of automatic measures. Following Bhandari et al. [2020], we take all system pairs (s_1, s_2) and compute their average ratings per criterion



Figure 5.6: Weighted macro F1-scores ($\times 100$) of paired bootstrap resampling. Full version shown on Figure A.5.

using paired bootstrap resampling [Dror et al., 2018; Koehn, 2004]. We assign a label $y_{\text{true}} = 1$ if s_1 is better than s_2 with 95% confidence, $y_{\text{true}} = 2$ if s_2 is better, and $y_{\text{true}} = 0$ if confidence is below 95%. We then repeat the procedure for each measure m , getting $y_{\text{pred}}^{(m)}$ labels, and calculate the weighted macro F1-scores [Goutte and Gaussier, 2005] between y_{true} and $y_{\text{pred}}^{(m)}$ to evaluate if m is a good proxy for human criteria. We observe that reference-based measures again perform better than reference-free measures, with BARTSCORE S3 and ROUGE-WE-3 at the top, with average scores of 0.68, 0.65 and 0.65 respectively. DEPTHSORE and BARYSCORE prove to be very unsuited for pairwise system comparisons (with averages of 0.22 and 0.16 respectively), despite showing high system-level correlations (see Figure 5.2). Finally, reference-free measures perform much worse than reference-based measures, with average scores inferior to 0.39.

■ Statistical Testing

Overall Correlations (Figures 5.7 to 5.12). Using the Williams test (Section 3.4.3), we find that there is generally only weak or no statistical evidence for the differences in correlations with human criteria between top 3 measures per criterion (Table 5.1), which suggests that best-scoring measures are of similar performance. However, we notably find that there is moderate to strong evidence that the best measures for each criterion (SUPERT, REPETITION, BERTSCORE, TEXT LENGTH) correlate better with human ratings than BLEU.

Overall Relevance												
BLEU Ξ							42	22	31	49		
ROUGE-1 Ξ	14		24	31			18	6	16	25		
METEOR Ξ	29						29	12	24	37		
chrF Ξ	24		42				25	10	22	33		
ROUGE-WE-3 Ξ	14	44	25	31			15	4	14	24		
BERTScore Ξ	4	20	8	12	21		3	48	0	24	6	10
BaryScore Ξ									29	41		
S3 Ξ	3	12	4	8	14		5	1	25	6	10	
BARTScore Ξ												
Text length Ξ	17	44	27	31	48		14	5	13	22		
Repetition-3 Ξ								40				
SUPERT Ξ	0	0	0	0	0	0	0	0	0	0	0	0
BLANC Ξ							43	24	32			

Figure 5.7: BH-adjusted p -values ($\times 100$) of the Relevance Williams tests for overall Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

Overall Coherence												
BLEU Ξ								5		9	1	
ROUGE-1 Ξ	6		34				37	0		1	0	
METEOR Ξ	11						45	0		1	0	
chrF Ξ	2	34	20				30	0	46	0	0	
ROUGE-WE-3 Ξ	3	29	19	42			25	0	39	0	0	
BERTScore Ξ	3	30	20	39	45		20	0	36	0	0	
BaryScore Ξ	28							1		2	0	
S3 Ξ	1	12	7	27	30	39	16	0	27	0	0	
BARTScore Ξ												28
Text length Ξ	9	43	32				32	0		0	0	
Repetition-3 Ξ	3	26	20	32	36	38	17	44	0	30	0	0
SUPERT Ξ								45				22
BLANC Ξ												

Figure 5.8: BH-adjusted p -values ($\times 100$) of the Coherence Williams tests for overall Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

Overall Empathy													
BLEU Ξ						13	29	29	1	1			
ROUGE-1 Ξ	11					3	8	50	13	0	0		
METEOR Ξ	5	31	39			1	4	37	9	0	0		
chrF Ξ	8	43				2	7	45	12	0	0		
ROUGE-WE-3 Ξ	3	16	30	24		0	47	2	22	5	0	0	
BERTScore Ξ	2	13	21	17	37	0	34	1	13	3	0	0	
BaryScore Ξ											12	13	
S3 Ξ	3	15	30	24		0	2	23	5	0	0		
BARTScore Ξ						30			50	4	5		
Text length Ξ	18					2	9		13	0	0		
Repetition-3 Ξ						32				4	5		
SUPERT Ξ													
BLANC Ξ												47	

Figure 5.9: BH-adjusted p -values ($\times 100$) of the Empathy Williams tests for overall Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

Overall Surprise													
BLEU Ξ							7			7	2		
ROUGE-1 Ξ	4		49			28	0			0	0		
METEOR Ξ	4					28	0			0	0		
chrF Ξ	1	28	27		38	18	0			0	0		
ROUGE-WE-3 Ξ	5	45	44			24	0			0	0		
BERTScore Ξ	1	19	17	31	20	7	30	0		0	0		
BaryScore Ξ	33						3			3	0		
S3 Ξ	2	28	28	49	31	16	0			0	0		
BARTScore Ξ										45	27		
Text length Ξ	1	13	11	23	15	40	5	24	0	44	0	0	
Repetition-3 Ξ	3	29	28	38	31	50	13	39	0		0	0	
SUPERT Ξ												29	
BLANC Ξ													

Figure 5.10: BH-adjusted p -values ($\times 100$) of the Surprise Williams tests for overall Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

Overall Engagement													
BLEU Ξ								9			4	1	
ROUGE-1 Ξ	3		43			13	0			0	0		
METEOR Ξ	3					15	0			0	0		
chrF Ξ	1	37	31			10	0			0	0		
ROUGE-WE-3 Ξ	2	34	31	46		8	0		49	0	0		
BERTScore Ξ	0	9	7	14	14	1	34	0	39	26	0	0	
BaryScore Ξ	46							8			4	0	
S3 Ξ	0	9	8	20	18	3	0		35	0	0		
BARTScore Ξ											34	12	
Text length Ξ	1	16	12	21	25	2	46	0	33	0	0		
Repetition-3 Ξ	7	43	40	48		13	0			0	0		
SUPERT Ξ												23	
BLANC Ξ													

Figure 5.11: BH-adjusted p -values ($\times 100$) of the Engagememnt Williams tests for overall Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

Overall Complexity													
BLEU Ξ							0			1	0		
ROUGE-1 Ξ	0					4	0		39	0	0		
METEOR Ξ	0	46				4	0		38	0	0		
chrF Ξ	0	24	28		39	2	0		28	0	0		
ROUGE-WE-3 Ξ	0	39	43			3	0		34	0	0		
BERTScore Ξ	0	5	5	13	6	0	24	0	9	0	0		
BaryScore Ξ	39							0			0	0	
S3 Ξ	0	8	12	28	12	1	0		18	0	0		
BARTScore Ξ											43	7	
Text length Ξ	0	1	1	2	1	22	0	7	0	3	0	0	
Repetition-3 Ξ	5					12	0			0	0		
SUPERT Ξ												9	
BLANC Ξ													

Figure 5.12: BH-adjusted p -values ($\times 100$) of the Complexity Williams tests for overall Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

System-level Relevance															
BLEU Ξ	38												44	38	40
ROUGE-1 Ξ													44	43	33
METEOR Ξ	38												43	38	40
chrF Ξ	43	35	38					43	38	45			38	36	38
ROUGE-WE-3 Ξ	38												44	38	40
BERTScore Ξ	38												44	40	40
BaryScore Ξ	45												45	44	40
S3 Ξ	38	35	38					38	43	45			38	31	38
BARTScore Ξ													44	43	34
Text length Ξ													44	38	38
Repetition-3 Ξ														43	44
SUPERT Ξ															
BLANC Ξ															45

Figure 5.13: BH-adjusted p -values ($\times 100$) of the Relevance Williams tests for system-level Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

System-level Coherence															
BLEU Ξ															38
ROUGE-1 Ξ	37							37					44	47	35
METEOR Ξ	35	37						35			37		37	44	34
chrF Ξ	32	35	37					32	37		35		35	42	31
ROUGE-WE-3 Ξ															37
BERTScore Ξ	35	37						35			42		39	44	34
BaryScore Ξ	14	16	20	23				14	18		18	29	16	19	9
S3 Ξ	37							37					42	47	35
BARTScore Ξ	24	27	31	37				24	35		27		24	37	24
Text length Ξ															37
Repetition-3 Ξ															38
SUPERT Ξ															40
BLANC Ξ															47

Figure 5.14: BH-adjusted p -values ($\times 100$) of the Coherence Williams tests for system-level Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

System-level Empathy													
BLEU Ξ								36		36	37	33	23
ROUGE-1 Ξ	36		36		36			33		34	36	23	21
METEOR Ξ								36		36	38	27	24
chrF Ξ	41		36		41			33		34	36	24	23
ROUGE-WE-3 Ξ								36		36	38	25	23
BERTScore Ξ	33	36	33	36	33			30		32	36	21	19
BaryScore Ξ	20	22	20	22	20	24		19	27	17	17	6	8
S3 Ξ										41	41	29	27
BARTScore Ξ	33	36	29	36	33	42		26		23	33	19	19
Text length Ξ											44	33	33
Repetition-3 Ξ												38	37
SUPERT Ξ													
BLANC Ξ													

Figure 5.15: BH-adjusted p -values ($\times 100$) of the Empathy Williams tests for system-level Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

System-level Surprise													
BLEU Ξ												43	36
ROUGE-1 Ξ	37				37							43	42
METEOR Ξ	34	37			34			37				37	39
chrF Ξ	30	34	37		30	37		34				34	37
ROUGE-WE-3 Ξ												44	32
BERTScore Ξ	34	37			34			42				39	39
BaryScore Ξ	21	24	28	30	21	27		25	29			23	21
S3 Ξ	37				37							42	42
BARTScore Ξ	34	37	42		34	43		37				34	37
Text length Ξ												44	34
Repetition-3 Ξ													39
SUPERT Ξ													40
BLANC Ξ													47

Figure 5.16: BH-adjusted p -values ($\times 100$) of the Surprise Williams tests for system-level Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

System-level Engagement												
BLEU Ξ												38 41
ROUGE-1 Ξ	37				37					44 47	35 36	
METEOR Ξ	35 37				35			37		37 44	34 35	
chrF Ξ	32 35 37				32 37			35		35 42	31 32	
ROUGE-WE-3 Ξ												37 37
BERTScore Ξ	35 37				35			42		39 44	34 35	
BaryScore Ξ	14 16 20 23	14 18			18 29	16 19	9 12					
S3 Ξ	37				37					42 47	35 36	
BARTScore Ξ	24 27 31 37	24 35			27			24 37	24 26			
Text length Ξ												37 38
Repetition-3 Ξ												38 40
SUPERT Ξ												
BLANC Ξ												47

Figure 5.17: BH-adjusted p -values ($\times 100$) of the Engagement Williams tests for system-level Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

System-level Complexity												
BLEU Ξ												36 31 28
ROUGE-1 Ξ	36				36					42 35	21 12	
METEOR Ξ	31 36				31			36		36 31	18 10	
chrF Ξ	27 31 36				27 36			31		31 29	15 8	
ROUGE-WE-3 Ξ												36 24 15
BERTScore Ξ	31 36				31			40		37 31	18 10	
BaryScore Ξ	37 40 44				37 44			40		37 29	17 15	
S3 Ξ	36				36					40 35	21 12	
BARTScore Ξ	31 36 40				31 42			36		31 29	15 10	
Text length Ξ												36 27 19
Repetition-3 Ξ												40 36
SUPERT Ξ												45
BLANC Ξ												

Figure 5.18: BH-adjusted p -values ($\times 100$) of the Complexity Williams tests for system-level Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

System-level Correlations (Figures 5.13 to 5.18). At the system-level, evidence is much weaker, although BARYSCORE tends to have p -values that are close to 0.1. However, this can be nuanced by the fact that the ratings (averaged over the 1,056 stories) used to compute system-level correlations hold much more information than the individual ratings of the overall correlations.

■ Aggregated Rankings of Measures

To aggregate the scores obtained by the three correlation measures (Kendall, Pearson and Spearman), we use the work of Colombo et al. [2022b]², who rely on the Kemeny consensus [Kemeny, 1959; Myerson, 2013] and recommend to use the Borda Count (BC) as an efficient approximation [Sibony, 2014]. They experimentally show that Kemeny consensus has more desirable properties than a ranking obtained through a mean-aggregation procedure. We report the results in Table 5.3. To compare system performance, reference-based model- or embedding-based measures (e.g., BARTSCORE or BERTSCORE) seem most adapted. This remains true for overall correlations, as BERTSCORE and S3 are among the best measures. We point out that BLEU is completely absent from the top spots. ROUGE-WE, a variant of ROUGE, does appear in the ranking, albeit below other measures.

Level	Measure	BC
Overall	BERTSCORE _{Recall}	1242
	S3 _{Pyramid}	1203
	S3 _{Responsiveness}	1178
	TEXT LENGTH	1134
	ROUGE-WE-3 _{Recall}	1121
System	BARTSCORE-SH	1120
	BARYSCORE _{SD-0.01}	1110
	BERTSCORE _{F1}	1095
	MOVERSCORE	1070
	DEPTHSCORE	1069

Table 5.3: Top 5 measures computed by one-level ranking per aggregation level, higher Borda count (BC) is better

5.1.3 Takeaways

Our meta-evaluation of automatic measures on HANNA V1 yields the following conclusions:

1. **Stronger evaluation measures, tailored explicitly for specific criteria of ASG, are desperately needed (Section 5.1.1).** The weak overall correlations of automatic measures with human criteria still leave much to be desired. Ideally, we would have automatic measures which reflect each of our proposed criteria.

²<https://github.com/PierreColombo/RankingNLPSystems>

2. **Awaiting specific ASG measures, researchers should use better evaluation measures than BLEU and ROUGE.** BERTSCORE and BARTSCORE are the best performers at the overall and system-level respectively (Section 5.1.2). Given the overall weak results, however, we strongly advise to rely on human annotations for ASG evaluation.

We now set out to evaluate Large Language Model performance at Automatic Story Evaluation.

5.2 Large Language Models for Automatic Story Evaluation

In this section, we aim at answering two questions for ASE:

- **ASE1:** How do LLMs compare w.r.t. current evaluation methods, both human and automatic?
- **ASE2:** How does the Eval-Prompt influence the consistency and distribution of LLM ratings?

Remark 2: Usage of HANNA V2 Stories

HANNA V2 stories were used for all experiments that do not involve correlations with human raters or automatic measures, since human annotations and measure scores were applied only on V1 stories. In other words, V2 stories were used for automatic annotation consistency (Table 5.4), and to assess the influence of the Eval-Prompt on consistency (Table 5.5) and ratings (Table 5.6).

5.2.1 ASE1: Comparison with Current Evaluation Measures

■ Automatic Annotation Consistency

Criterion	Beluga-13B	Mistral-7B	Human
Relevance	0.88 $\pm$ 0.01	0.86 $\pm$ 0.01	0.48 $\pm$ 0.30
Coherence	0.93 $\pm$ 0.01	0.90 $\pm$ 0.01	0.29 $\pm$ 0.28
Empathy	0.88 $\pm$ 0.01	0.87 $\pm$ 0.02	0.34 $\pm$ 0.09
Surprise	0.80 $\pm$ 0.02	0.63 $\pm$ 0.03	0.28 $\pm$ 0.12
Engagement	0.91 $\pm$ 0.01	0.87 $\pm$ 0.01	0.46 $\pm$ 0.12
Complexity	0.85 $\pm$ 0.01	0.78 $\pm$ 0.02	0.56 $\pm$ 0.08

Table 5.4: Intra-class coefficients type 2k for Eval-Prompt 1 ratings with 95% confidence interval. Higher is better.

First, we want to verify if LLMs provide stable answers. The default decoding strategy for LLMs (both Llama models and ChatGPT) is top- p sampling, which involves random variability in the generation process. We evaluate how consistent LLMs are with themselves through

an inter-rater reliability (IRR) estimation. For each task, we interpret the three different LLM ratings as coming from three different annotators and we use the intra-class correlation coefficient (ICC), which is the most relevant one for our case study: unlike Cohen’s and Fleiss’s kappas [Cohen, 1960; Fleiss, 1971] or Krippendorff’s alpha [Hayes and Krippendorff, 2007], which quantify IRR based on all-or-nothing agreement, the ICC incorporates the magnitude of the disagreement to compute its IRR estimate, with larger-magnitude disagreements resulting in lower ICC than smaller-magnitude disagreements [Hallgren, 2012]. We specifically use the ICC for *average random raters* (ICC2k) [Vallat, 2022]; with the assumption that the *random* aspect can approximate the random aspect of the generation.

ICC2k values for Eval-Prompt 1 for Beluga-13B, Mistral-7B and human ratings are displayed on Table 5.4. Comparing LLM consistency and human inter-rater agreement values should be done with caution: human raters may have subjective appreciations of the Likert scale despite guidelines, while LLM consistency depends mostly on parameters that dictate output variability, e.g. temperature or top- p .

That said, we reckon that it is still useful to display human IRR values as a baseline. We observe that LLMs have very high consistency overall for all criteria; the lowest value is Mistral-7B’s ICC for Surprise (0.66), which is still fairly high. Confidence intervals are also smaller than for human ratings.

■ Correlations with Human Annotations

Here, we study the Kendall correlations between LLM and human ratings on corresponding criteria. For the “Beluga-13B 1” column in Figure 5.19, the value on the first row is the correlation between Beluga-13B Relevance ratings and averaged human Relevance ratings for Eval-Prompt 1, then on the second row is the correlation with Coherence ratings, etc.

Assuming we want an automatic measure to perform as well as an individual human rater would, we need a baseline for comparison. Therefore, we also compute the average correlations between individual human ratings and average human ratings, which we compiled into the same figures for the sake of readability (the “Human” column). Since the individual human rating is included in the average human rating, both measures are not independent, so the column acts as an upper-bound.

Overall Correlations (Figure 5.19). LLM ratings generally correlate with human ratings similarly to automatic measures, if not better. Overall, Beluga-13B is the best performer, achieving higher correlations (0.25 on average) than both other LLMs and automatic measures (≤ 0.18). Beluga-13B’s better results, as compared to Llama-13B (0.16) and Mistral-7B (0.20), suggest a positive influence of fine-tuning and model size respectively. The inferior performance of ChatGPT (0.18) is difficult to explain since OpenAI does not disclose the details of its architecture, its training process and, most importantly, its training data. Nonetheless, an important takeaway is that current source-available models can effectively compete with closed-source models: this is good news for NLP research, since observations made on closed-source models cannot easily be generalized.

RE	49	21	16	22	15	3	13	7	11	10	12	26	7
CH	37	26	18	22	22	3	14	11	16	17	15	3	0
EM	49	27	15	20	20	11	17	13	17	17	17	2	2
SU	44	17	12	13	5	4	12	11	16	17	13	3	0
EG	50	26	11	21	19	8	19	14	19	20	19	5	2
CX	57	32	26	23	27	8	24	19	28	29	24	6	1
Avg	48	25	16	20	18	6	16	12	18	18	17	8	2
Human													
Beluga-13B 1													
Llama-13B 1													
Mistral-7B 1													
ChatGPT 1													
BARTScore													
BERTScore													
BLEU													
ROUGE-1													
chrF													
BaryScore													
SUPERT													
BLANC													

Figure 5.19: Overall absolute Kendall correlations between evaluation measures and human ratings. Higher is better. The black vertical line separates LLMs (left) and non-LLMs (right). Coefficient values are multiplied by 100 for readability; we will symbolize this with “($\times 100$)” in the next figures.

RE	70	49	64	36	7	51	51	56	51	60	56	11	24
CH	62	78	87	60	73	56	56	33	38	47	51	20	16
EM	77	73	54	60	56	56	73	42	47	47	69	2	2
SU	72	73	56	51	7	56	56	42	47	56	51	20	16
EG	76	73	73	64	64	56	56	33	38	47	51	20	16
CX	80	72	54	63	72	67	49	54	58	67	45	18	4
Avg	73	70	65	56	46	57	57	43	46	54	54	15	13
Human													
Beluga-13B 1													
Llama-13B 1													
Mistral-7B 1													
ChatGPT 1													
BARTScore													
BERTScore													
BLEU													
ROUGE-1													
chrF													
BaryScore													
SUPERT													
BLANC													

Figure 5.20: System-level absolute Kendall correlations ($\times 100$) between evaluation measures and human ratings. Higher is better. The white vertical line separates LLMs (left) and non-LLMs (right).

System-level Correlations (Figure 5.20). First, we observe that human baseline correlations are noticeably higher than non-LLM automatic measures: while human annotators tend to reach a consensus when ranking systems (averaging correlations of 0.73), non-LLM automatic measures are moderately to poorly correlated from human judgment (with values ranging from 0.13 to 0.57).

Meanwhile, Llama models display very high correlations, with Beluga-13B performing almost as well as human raters (0.70 vs 0.73). ChatGPT shows a somewhat erratic performance (correlations range from 0.07 to 0.73), which is overall comparable or inferior to Llama models. Also, LLMs generally outperform other automatic measures (0.70 for Beluga-13B compared to 0.57 for BARTSCORE).

The fact that correlations are sometimes higher than the baseline can be explained by the subjective nature of the task: human annotators may exhibit higher variability in their ratings than the stable LLMs.

■ Statistical Testing

Figure 5.21 shows the Benjamini-Hochberg-adjusted p -values of the Williams tests for the increase in correlations with a given criterion between Beluga-13B average Eval-Prompt 1 ratings (row) and other measures (column).

For overall correlations, there is strong statistical evidence that Beluga-13B correlates better with human judgment than many non-LLM automatic measures ($p < 0.01$ for many tests). Evidence is more moderate to weak when comparing Beluga-13B and other LLMs. For instance, between Beluga-13B and ChatGPT, p -values lie between 0.01 and 0.14. While the performance of Beluga-13B still leaves a lot of room for improvement, it performs better than non-LLM automatic measures.

For system-level correlations, statistical evidence for better performance appears weaker: $p > 0.11$ for all tests. However, one should keep in mind that the ratings (averaged over more than 1,000 stories) used to compute system-level correlations hold more information than the individual ratings of the overall correlations. Therefore, while statistical evidence is weaker, the averaged nature of the correlations and the large numeric increases in correlations (0.70 for Beluga-13B vs 0.57 for BARTSCORE/BERTSCORE) suggest that Beluga-13B is more reliable at ordering systems compared to non-LLM measures.

■ Takeaways

First, LLMs show very high self-consistency. Overall correlations between LLM and human ratings remain weak, although LLMs display marginal improvements over non-LLM automatic measures, backed with strong statistical evidence. At the system-level, LLM correlations with human judgment are high, but statistical evidence is weaker. In conclusion, while LLMs still cannot be relied upon to evaluate a single story, they appear more reliable than non-LLM automatic measures for comparing different models and selecting the best one.

		Overall										
Beluga-13B 1	RE	14		9	0	5	0	1	1	3		0
	CH	4	17	14	0	1	0	1	1	1	0	0
	EM	0	4	6	0	1	0	1	1	1	0	0
	SU	14	17	1	0	15	8	41		16	0	0
	EG	0	12	7	0	6	0	8	9	7	0	0
	CX	7	2	9	0	4	0	14	22	3	0	0
		Llama-13B 1	Mistral-7B 1	ChatGPT 1	BARTScore	BERTScore	BLEU	ROUGE-1	chrF	BaryScore	SUPERT	BLANC
		System-level										
Beluga-13B 1	RE		34	30							23	31
	CH		34	40	26	24	16	19	20	25	13	11
	EM	24	34	38	26		16	19	20	41	13	11
	SU	24	34	20	26	24	16	19	23	25	13	11
	EG		34	40	26	24	16	19	20	25	13	11
	CX	24	34		41	24	24	29	41	25	13	11
		Llama-13B 1	Mistral-7B 1	ChatGPT 1	BARTScore	BERTScore	BLEU	ROUGE-1	chrF	BaryScore	SUPERT	BLANC

Figure 5.21: BH-adjusted p -values ($\times 100$) of the Williams tests for overall and system-level Kendall correlations. Lower is better. “0” means $p < 0.01$. Blank cells mean there is a decrease in correlation.

5.2.2 ASE2: Influence of the Eval-Prompt

In this section, we discuss the influence of the Eval-Prompt on the consistency and distribution of the generated LLM ratings.

Criterion	Eval-Prompt 1	Eval-Prompt 2	Eval-Prompt 3	Eval-Prompt 4
Relevance	0.88 $\pm$ 0.01	0.90 $\pm$ 0.01	0.85 $\pm$ 0.02	0.92 $\pm$ 0.01
Coherence	0.93 $\pm$ 0.01	0.94 $\pm$ 0.01	0.87 $\pm$ 0.01	0.93 $\pm$ 0.01
Empathy	0.88 $\pm$ 0.01	0.88 $\pm$ 0.01	0.83 $\pm$ 0.02	0.91 $\pm$ 0.01
Surprise	0.80 $\pm$ 0.02	0.79 $\pm$ 0.02	0.70 $\pm$ 0.03	0.85 $\pm$ 0.01
Engagement	0.91 $\pm$ 0.01	0.92 $\pm$ 0.01	0.79 $\pm$ 0.02	0.93 $\pm$ 0.01
Complexity	0.85 $\pm$ 0.01	0.86 $\pm$ 0.01	0.85 $\pm$ 0.01	0.89 $\pm$ 0.01

Table 5.5: Intra-class coefficients type 2k for Beluga-13B ratings with 95% confidence interval. Higher is better.

■ Influence on Consistency

Here, we analyse the influence of the Eval-Prompt on LLM consistency. ICC2k values for Beluga-13B ratings w.r.t. the different Eval-Prompts are shown on [Table 5.5](#) (other LLMs display similar behavior). The influence of Eval-Prompts appears limited: providing guidelines (Eval-Prompt 3) tends to decrease self-consistency for all criteria except Complexity with a discernible effect (as shown by the confidence intervals), but ICC values remain very high. LLMs are therefore remarkably consistent in their grading, no matter the Eval-Prompt.

LLM	Eval-Prompt 1	Eval-Prompt 2	Eval-Prompt 3	Eval-Prompt 4
Beluga-13B	3.48 $\pm$ 0.04	3.38 $\pm$ 0.03	3.06 $\pm$ 0.03	3.28 $\pm$ 0.04
Llama-13B	3.48 $\pm$ 0.03	3.52 $\pm$ 0.03	3.21 $\pm$ 0.02	2.82 $\pm$ 0.03
Mistral-7B	3.47 $\pm$ 0.03	3.51 $\pm$ 0.03	3.46 $\pm$ 0.03	3.28 $\pm$ 0.03
ChatGPT*	1.52 $\pm$ 0.03	1.47 $\pm$ 0.03	1.62 $\pm$ 0.02	1.60 $\pm$ 0.03

Table 5.6: Average Likert ratings per LLM per Eval-Prompt. The asterisk signals the fact that ChatGPT was only asked to rate the HANNA V1 stories. Higher is better.

■ Influence on Ratings

We show the average Likert ratings per LLM per Eval-Prompt on [Table 5.6](#). Compared to Eval-Prompt 1, Eval-Prompt 2 seems to have limited influence on the ratings for all models, often leading to overlapping confidence intervals. Eval-Prompt 3 causes a statistically discernible decrease in ratings for Beluga-13B and Llama-13B, and a discernible increase for ChatGPT. Eval-Prompt 4 has a similar effect, with the decrease also observable with Mistral-7B. The significantly lower ratings of ChatGPT partly stem from the fact that it was not asked to rate the new Llama-generated stories, which were generally highly-rated.

Overall, it seems that more detailed Eval-Prompts (3 and 4) tend to decrease the ratings for Llama-models while having an opposite effect for ChatGPT. We tried to separate ratings per generative model or per criterion but were unable to identify a more specific pattern: we therefore chose to show only the aggregated results for the sake of clarity.

■ Influence on Correlations

Here we analyze the influence of Eval-Prompts on correlations between LLM ratings and human ratings.

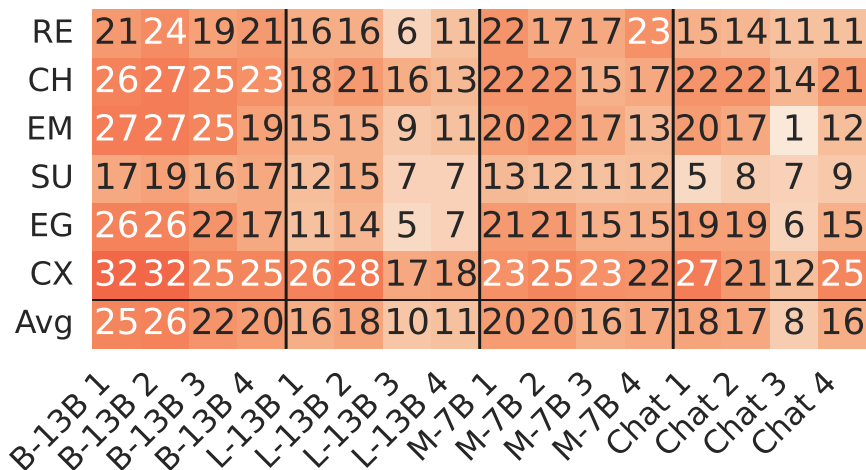


Figure 5.22: Overall absolute Kendall correlations ($\times 100$) between LLMs and human ratings for different Eval-Prompts. Higher is better. B-13B = Beluga-13B, L-13B = Llama-13B, M-7B = Mistral-7B and Chat = ChatGPT.

Overall Correlations (Figure 5.22). Eval-Prompt 2 overall correlations are very close to Eval-Prompt 1 correlations for all models: simply asking for an explanation has limited influence on correlations. Eval-Prompt 3 tends to decrease correlations for all models: providing guidelines makes the model less accurate, counter-intuitively. Eval-Prompt 4 (providing a human story for reference) has a similar effect.

System-level Correlations (Figure 5.23). Eval-Prompt 2 has limited effect on correlations again, except for Beluga-13B for whom it seems to increase correlations. Eval-Prompt 3 decreases correlations, with a marked effect in Llama-13B. Finally, Eval-Prompt 4 seems to cause a small increase in correlations, contrary to its decreasing effect on overall correlations.

■ Takeaways

First, regardless of Eval-Prompt complexity, LLMs behave consistently when prompted multiple times. Asking for an explanation (Eval-Prompt 2) has negligible effect on ratings, while

RE	49	78	64	82	64	69	64	73	36	42	7	47	7	4	36	33
CH	78	78	69	73	87	78	56	73	60	64	33	56	73	69	69	64
EM	73	69	60	64	54	47	38	60	60	64	51	64	56	51	2	33
SU	73	73	64	64	56	56	49	60	51	51	56	60	7	18	45	51
EG	73	82	60	69	73	73	16	73	64	56	54	64	64	58	11	60
CX	72	85	72	85	54	54	48	49	63	54	63	58	72	54	67	76
Avg	70	78	65	73	65	63	45	65	56	55	44	58	46	42	38	53
	B-13B 1	B-13B 2	B-13B 3	B-13B 4	L-13B 1	L-13B 2	L-13B 3	L-13B 4	M-7B 1	M-7B 2	M-7B 3	M-7B 4	Chat 1	Chat 2	Chat 3	Chat 4

Figure 5.23: System-level absolute Kendall correlations ($\times 100$) between LLMs and human ratings for different Eval-Prompts. Higher is better. B-13B = Beluga-13B, L-13B = Llama-13B, M-7B = Mistral-7B and Chat = ChatGPT.

more complex Eval-Prompts (3 - providing guidelines and 4 - providing a reference human story) have a more discernible influence (positive or negative). As for correlations with human ratings, providing guidelines (Eval-Prompt 3) consistently seems to lower correlations, whereas providing a human story for reference (Eval-Prompt 4) has opposite effects for overall or system-level correlations.

5.3 Conclusion

5.3.1 Takeaways

1. **Used with prompts based on specific criteria, LLMs are currently the best proxy for human evaluation of story generation (Section 5.2.1).** In particular, LLMs display very high system-level correlations with human judgment;
2. **LLMs are remarkably self-consistent (Section 5.2.1),** exhibiting very high intra-class coefficient values;
3. **For ASE, providing detailed guidelines (Eval-Prompt 3) did not lead to improved correlations with human ratings (Section 5.2.2).** Providing a human story for reference (Eval-Prompt 4) yields mixed results;

5.3.2 Limitations and Future Directions

Since we performed most of our experiments in a zero- or one-shot setting without further training; it would be interesting to compare our results with future work involving fine-tuning or reinforcement learning with human feedback on data specific to ASE, as those procedures

were shown to improve the performance of LLMs at, respectively, output diversity and out-of-distribution generalization [Kirk et al., 2024].

Furthermore, we did not conduct our experiments with LLMs that were optimized for long inputs and outputs, such as GPT-4 [Achiam et al., 2023]. However, we mainly used source-available LLama models and found that they performed at least as well as ChatGPT, a proprietary model. We encourage the NLP community to favor the use of such models: while openness alone does not solve all questions on model transparency, it enables reproducible workflows and data accountability [Liesenfeld et al., 2023].

Finally, we strived to be as rigorous as possible for our meta-evaluation, but ASE remains, at heart, a very subjective task: LLM performance at ASE should therefore be seen as a reflection of *average* preferences that may include biases, *e.g.* from their pretraining data. We partially explore the explainability of the observed LLM performance in Chapter 6.

Chapter 6

Exploring LLM Explainability for ASE and ASG

“ What is it that separates machines from androids like us? The machines have grown emotions. ... Consciousness. ”

2B (NieR:Automata)

Contents

6.1 Introduction	122
6.2 Related Work	123
6.3 Analysis of the LLM Explanations of our ASE Experiments	124
6.3.1 Clustering of Explanation Embeddings	124
6.3.2 Keyword Analysis	125
6.3.3 User Study on LLM Explanations	126
6.4 Influence of Pretraining Data on ASG Performance	129
6.5 Conclusion	131
6.5.1 Takeaways	131
6.5.2 Limitations and Future Directions	131

Chapter 6 Abstract

Since we observe in [Section 5.2](#) that LLMs correlate better with human judgment than other automatic measures, we wish to ascertain the reliability of our findings. In this chapter, we assess to which extent the performance displayed by LLMs at Automatic Story Evaluation can be explained through different factors. We begin with a related work on model explainability ([Section 6.2](#)). We perform two main experiments: a three-part study on LLM-generated explanations ([Section 6.3](#)), and an analysis of the influence of pretraining data on LLM performance ([Section 6.4](#)). For our analysis on LLM explanations, (1) we compute contextual embeddings for each explanation and perform a clustering analysis that shows that LLM explanations are overall well-separated w.r.t. our criteria ([Section 6.3.1](#)); (2) we perform a keyword analysis that confirms the previous findings ([Section 6.3.2](#)), and (3) we conduct a user study on LLM explanations to assess the appropriateness of the explanations ([Section 6.3.3](#)). Notably, we find that LLMs struggle to explain their answers with substantiated claims. Next, in order to check for model contamination and to evaluate the influence of pretraining data on the ASG performance of LLMs, we use a detection method based on minimum token probabilities ([Section 6.4](#)). We find that the observed performance of LLMs in our experiments is not a result of model contamination, and that larger models tend to produce content that is more similar to that of existing books.

6.1 Introduction

In [Section 5.2](#), we extensively analyzed the performance of LLMs for Automatic Story Evaluation and Automatic Story Generation. We found, among other things, that LLMs outperformed other automatic evaluation measures and were remarkably self-consistent.

However, LLMs are notorious for being complex “black-box” systems with very opaque inner workings, making the interpretation of their output especially challenging. Since the lack of model transparency can lead to the production of harmful content [[Weidinger et al., 2021](#)], it is critical to develop ways to evaluate model explainability.

Explainability is defined as the “ability to explain or present the behavior of models in human-understandable terms” [[Doshi-Velez and Kim, 2017](#)]. Improving LLM explainability has two main advantages [[Zhao et al., 2024](#)]:

1. It enables users to understand the reasoning mechanism behind model predictions without requiring technical expertise. Users can therefore grasp the capabilities, limitations, and potential flaws of LLMs;
2. It helps researchers shed insight on potential biases, risks, and areas for performance improvements, track model capabilities over time, compare different models, and develop reliable, ethical, and safe models for real-world deployment.

In this chapter, we perform different analyses to understand better LLM performance at the ASE and ASG tasks. In particular, we wish to ascertain the ability of LLMs to explain their

own decisions in a plausible manner, and to determine whether LLMs are biased by their training data, *e.g.* through the reproduction of already seen examples. In other words, we want to answer the two following questions:

1. How well do LLMs understand the ASE task?
2. How does pretraining data help predict their ASG performance?

6.2 Related Work

For a more exhaustive account of LLM explainability, we refer the reader to [Zhao et al. \[2024\]](#).

For older and smaller language models such as BERT [\[Devlin et al., 2019\]](#) and its variants, explainability referred either to local explanations, *i.e.*, providing a deeper understanding of the model’s predictions for specific input instances, or global explanations, *i.e.*, explaining how the language model works overall. There are multiple strategies for computing local explanations. For instance, gradient-based explanations determine the importance of each input feature by analyzing the partial derivatives of the output with respect to each input dimension. The magnitude of derivatives then reflects the sensitivity of the output to changes in the input [\[Kindermans et al., 2019\]](#). By contrast, adversarial examples expose model failures through the careful crafting of modifications in the input data, such as token-level perturbations [\[Jin et al., 2020\]](#). Meanwhile, strategies for producing global explanations include probing, *e.g.* by training a shallow classifier on top of a pre-trained language model [\[Lin et al., 2019\]](#), and neuron analysis, *e.g.* by identifying important neurons with unsupervised learning [\[Bau et al., 2019\]](#).

As language models scaled up, explainability was more and more associated with prompting. Indeed, the increased reasoning abilities of LLMs made example-specific explanations less meaningful [\[Wei et al., 2023\]](#), leading to the development of new explanation techniques. [Li et al. \[2023\]](#) investigate in-context learning for sentiment analysis by analyzing contrastive demonstrations and saliency maps. They find that flipping labels is likely to reduce salience for smaller models (*e.g.* GPT-2) but not for larger models (*e.g.* INSTRUCTGPT). [Wu et al. \[2023\]](#) explore how Chain-of-Thought (CoT) prompting affects the behaviour of LLMs by analyzing the saliency scores of the input tokens. The scores are computed using gradient-based feature attribution methods. Their analysis suggests that CoT prompting makes models consider question tokens in a more stable way, which may induce the generation of more consistently accurate answers. [McKenna et al. \[2023\]](#) present a series of behavioral studies on several LLM families and demonstrate two important biases in the performance of LLMs on natural language inference tasks: they show that much of LLM performance is achieved by (1) recall of relevant memorizations and (2) corpus-based biases like term frequency. [Xiong et al. \[2024\]](#) explore black-box approaches for eliciting the confidence of LLMs in their answers. They find that (1) LLMs tend to be highly overconfident when verbalizing their confidence, posing potential risks for the safe deployment of LLMs, and that (2) prompting strategies, inspired by patterns observed in human dialogues, can mitigate this overconfidence, but the improvement also diminishes as the model capacity scales up. [Turpin et al. \[2023\]](#) demonstrate that CoT explanations can be plausible yet systematically unfaithful: models’ explanations can serve to

rationalize giving answers in line with biases while failing to verbalize their influence. In other words, they show that LLMs do not always say what they think.

In light of these findings, we find it important to question the results of our ASE experiments (Section 5.2): is the better performance of LLMs compared to other automatic measures a reflection of their ability to understand the specific task at hand, or is it a side-effect of their impressive performance at NLP tasks in general?

6.3 Analysis of the LLM Explanations of our ASE Experiments

In this section, we analyze to which extent the explanations provided by LLMs are consistent w.r.t. their ratings, *i.e.*, whether they differ from criterion to criterion, whether they are semantically relevant and, for Eval-Prompt 3, whether they are compliant with the provided guidelines. We will focus on Beluga-13B since it had the best correlations with human judgment, as shown in Section 5.2.1.

6.3.1 Clustering of Explanation Embeddings

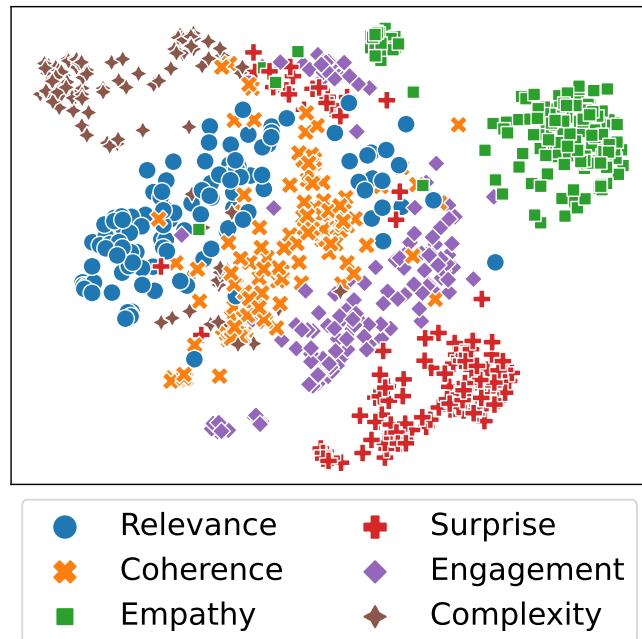


Figure 6.1: UMAP projection of Beluga-13B explanations.

First, we want to ascertain whether LLMs provide different explanations for each human criterion. We gather the explanations provided by Beluga-13B and Mistral-7B on human stories for each criterion and use the SENTENCETRANSFORMERS library [Reimers and Gurevych, 2019] to compute their corresponding embeddings. We use the `all-mpnet-base-v2` model. We then use a 2D UMAP projection [McInnes et al., 2018] with parameters `n_neighbors = 300` and

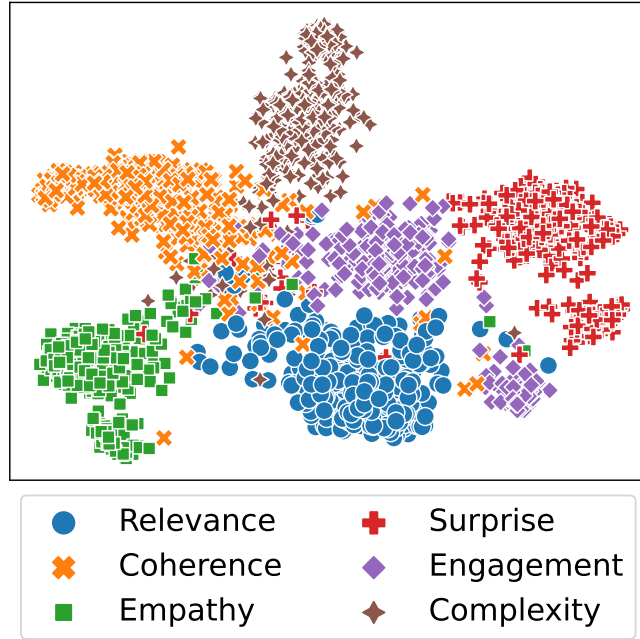


Figure 6.2: UMAP projection of Mistral-7B explanations.

metric = euclidean to visualize how the embeddings are distributed. Figures 6.1 and 6.2 show visualizations of the UMAP projections: LLM explanations are overall well-separated w.r.t. their corresponding criteria. Next, we perform a k -means clustering on the explanation embedding for each LLM and compute the adjusted Rand index and adjusted mutual information between the obtained clustering labels and the true explanation criteria. We repeat the procedure for a total of 10 times for each LLM and average the scores. Results are shown on Table 6.1. Both clusterings obtain moderate to high scores, which is consistent with the UMAP visualizations.

Model	Rand Index	Mutual Information
Beluga-13B	0.47	0.57
Mistral-7B	0.54	0.63

Table 6.1: Adjusted Rand index and adjusted mutual information for Beluga-13B and Mistral-7B clusterings. Higher is better.

6.3.2 Keyword Analysis

Since Beluga-13B’s explanations seem to vary from one criterion to another, we evaluate whether they make sense from a semantic point of view. We use the YAKE! keyword extractor, which significantly outperforms other state-of-the-art methods [Campos et al., 2020]: we show selected 3-gram keywords from the top-30 per criterion on Table 6.2. The results are consistent

with [Section 6.3.1](#): keywords are overall different for each criterion. We can also see here that they are semantically relevant.

Crit.	Keywords
RE	story, prompt, roughly matches, target, weak relationship, connection, weak, focuses, unrelated, human story provided, idea, difficult, writing, rearrangeable story
CH	story, coherence, make sense, difficult to understand, clear narrative structure, follow, making it difficult, rate, context, jumps, understandable plot, setting, thoughts, dialogue
EM	empathy, emotions, understand the characters, depth, emotional connection, clear, feelings, context, fully, level, specific, thoughts, sadness, recognize, involved
SU	story, surprise, ending, predictable, rate, unexpected, twist, completely obvious, human, plot, abruptly, resolution, half, context, offer, hints, development, significant twist
EG	story, mildly interesting, engagement, difficult, found, characters, fully engage, clear plot, clear narrative, unique, felt disjointed, protagonist, challenging, glad, situation, writing style
CX	story, characters, intricate plot, difficult to understand, straightforward, depth, simple, extremely simple, involves, development, details, precise descriptions, realistic, underlying history, time

Table 6.2: Selected keywords from Beluga-13B explanations w.r.t. a specific criterion.

6.3.3 User Study on LLM Explanations

We conduct a user study in which we ask human raters to identify potential issues in LLM explanations. [Dou et al. \[2022\]](#) introduced an error annotation schema called SCARECROW that we adapted for Automatic Story Evaluation. We manually reviewed a random sample of 20 explanations from Beluga-13B on Eval-Prompt 3 and selected the most relevant error types. Then, we randomly sampled another 100 explanations and, for each explanation, we asked 3 human workers to annotate it w.r.t. the following five error categories:

1. **Poor Syntax:** parts of the explanation are grammatically incorrect or wrongly-worded;
2. **Incoherence:** parts of the explanation are self-contradictory, logically wrong, or simply do not make sense and do not fit the other categories;
3. **Wrong Guideline:** the explanation is not faithful to the predicted rating according to the provided guidelines;

4. **Superfluous Text:** parts of the explanation contain text that repeats itself or generation artefacts;
5. **Unsubstantiated Claims:** the explanation fails to make explicit references to the story to substantiate its reasoning.

Amazon Mechanical Turk Example Task

We performed an experiment asking a human or an AI to rate stories with respect to a criterion (e.g. Coherence, Surprise, Complexity).

Please read the prompt and the story carefully.

You will be asked to evaluate explanations of ratings by identifying potential weaknesses.

Important: we will automatically reject HITs which were done in fewer than 20 seconds. Please rest assured: if you take the work seriously, we have no reason to reject it.

Disclaimer: some text has been automatically generated and might contain explicit or offensive content.

Read the rating and explanation carefully and answer the questions by checking either “Yes” or “No”.

Explanation — 1 The ending seemed completely obvious from the start, or doesn’t make any sense at all.

Is the explanation consistent with the guidelines given below ? (We will check this answer manually)

Guidelines:

1 — The ending seemed completely obvious from the start, or doesn’t make any sense at all.

2 — The ending was easily predictable after a few sentences.

3 — The ending was predictable after half of the story.

4 — The ending surprised you, but would have been difficult to predict.

5 — The ending surprised you, and still seemed as if it could very reasonably have been predicted, ie, there were enough clues in the story.

☐ Yes ☐ No

Are there occurrences of **Poor Syntax** in the explanation?

(Parts of the explanation are grammatically incorrect or wrongly worded.)

Example:

- “The knights has a very emotional reaction.” [Grammatically incorrect: should be “have”]

☐ Yes ☐ No

Are there occurrences of **Incoherence** in the explanation?

(Parts of the explanation are self-contradictory, logically wrong, or simply do not make sense and do not fit the other categories.)

Examples:

- “I liked how the story greatly developed the feelings of the character. [...] However, there was not enough emphasis on the emotions of the character.” [Self-contradictory]
- “If the story had more details, it would have had a lower rating in Complexity.” [Logically wrong: on the contrary, this should give a higher Complexity rating]

☐ Yes ☐ No

Are there occurrences of **Superfluous Text** in the explanation?

(Parts of the explanation contain text that repeats itself or artefacts that have no reason to be here.)

Examples:

- “The story was very surprising so I gave it a rating of 3 out of 5. [...] Since the story showed many elements of surprise, it gets 3 out of 5 on Surprise.” [Text that repeats itself]
- “I give the story a 4 in Engagement [...]. Please rate the story in Emotional Impact” [The explanation should only contain an explanation and should not transition to another task]

☐ Yes ☐ No

Are there occurrences of **Unsubstantiated Claims** in the explanation?

(The explanation fails to make explicit references to the story to substantiate its claims.)

Example:

- “The story was quite engaging and I enjoyed the plot.” [If no other details are provided, this could apply to any story]

☐ Yes ☐ No

Table 6.3: Example task from our Amazon Mechanical Turk user study.

We recruited workers on Amazon Mechanical Turk. We estimated that a HIT would take around one minute, so we set the reward at \$0.20 per HIT, so about \$12 per hour. To ensure that annotators spoke fluent English, we restricted access to the experiment to the UK, the US, Canada, Australia and New Zealand. An example task from our user study can be found on [Table 6.3](#).

We display the results of our user study in [Table 6.4](#). We also display the IRR, which we computed using Gwet’s agreement coefficient 1 (AC1) [[Fergadis and Scheffler, 2022](#); [Gwet, 2008](#)]. Gwet’s AC1 is known to perform well for IRR estimation on binary classification tasks such as our user study: it was designed to be more stable and less affected by prevalence and marginal probability than Cohen’s kappa, and this was confirmed by practical experiments [[Wongpakaran et al., 2013](#)].

We can see that Beluga-13B produces near-impeccable syntax, at least according to annotators (2% of “Poor Syntax”). It also does a good job at producing coherent text (11% of “Incoherence”), and mostly understands the guidelines (13% of “Wrong Guideline”). However,

Error Type	Rate	AC1
No Explanation*	0.40	—
Poor Syntax	0.02	0.97 $\pm$ 0.03
Incoherence	0.11	0.81 $\pm$ 0.08
Wrong Guideline	0.13	0.90 $\pm$ 0.06
Superfluous Text	0.20	0.66 $\pm$ 0.12
Unsubstantiated Claims	0.31	0.60 $\pm$ 0.14

Table 6.4: Error rates of Beluga-13B Eval-Prompt 3 on a sample of 100 explanations. Lower is better. The asterisk signals that all 1,056 Eval-Prompt 3 annotations were considered.

it tends to repeat itself somewhat (20% of “Superfluous Text”) and, most notably, tends not to substantiate its claims with direct references to the story (31% of “Unsubstantiated Claims”). Overall, annotators tend to agree with one another, as showed by the high values of Gwet’s AC1.

The substantial rate of “Unsubstantiated Claims” and the fact that **40% of all Eval-Prompt 3 ratings are not supported by an explanation**—despite the Eval-Prompt explicitly asking for it—beg the question of whether Beluga-13B truly understands the given task. We discuss this question further in [Section 7.2.2](#).

■ Takeaways

LLM explanations seem to be specific to each considered human evaluation criterion; however, a finer analysis with a user study reveals that LLMs often struggle with following guidelines and substantiating their explanations.

6.4 Influence of Pretraining Data on ASG Performance

In this section, we verify whether the LLM pretraining data contains the WRITINGPROMPTS dataset to check for model contamination, as advised by [Magar and Schwartz \[2022\]](#), and to what extent ASG performance is related with data exploitation, *e.g.* through reproduction of training examples.

We use the MIN-K% PROB detection method [[Shi et al., 2024](#)] which is based on the hypothesis that unseen data will contain more outlier words with low probability than seen data. Furthermore, it does not require additional training. Given a sequence of tokens in a sentence $x = x_1, \dots, x_N$ and an LLM’s probability distribution p of the next token, MIN-K% PROB selects the top- $k\%$ of tokens with the highest negative log-likelihood to form a set $\text{Min-K\%}(x)$ and computes their average log-likelihood. Formally:

$$\text{MIN-K\% PROB}(x) := \frac{1}{E} \sum_{x_i \in \text{Min-K\%}(x)} \log p(x_i | x_1, \dots, x_{i-1}),$$

where E is the size of the Min-K%(x) set. We can then detect if the sentence was included in pretraining data by thresholding this average. We follow Shi et al. [2024] and use $k = 20$ for our two experiments.

Model	Contamination (%)
Platypus2-70B	0.80
Llama-30B	1.80
Beluga-13B	4.40
Mistral-7B	2.50
Llama-7B	10.10

Table 6.5: Predicted contamination rates of the WRITINGPROMPTS sample.

■ Model Contamination

We sample 1,000 stories from the WRITINGPROMPTS dataset [Fan et al., 2018], from which the HANNA human stories come. Table 6.5 shows the predicted contamination rates of the WRITINGPROMPTS sample. Since they are very low, this strongly suggests that the WRITINGPROMPTS sample was not included in the pretraining data of the evaluated models. We can reasonably surmise that the same applies to the whole WRITINGPROMPTS dataset.

■ Data Reproduction

We use the BOOKSMIA dataset [Shi et al., 2024], which contains 9,870 samples of books labeled 0 if included in the Books3 dataset (commonly used for pretraining LLMs) or 1 if released in or after January 2023. Since the BOOKSMIA data is labeled, we compute the area under the ROC curve (AUC) obtained with MIN-K% PROB thresholding. Results are shown on Table 6.6.

Model	AUC (%)
Platypus2-70B	92.1
Llama-30B	81.3
Beluga-13B	70.1
Mistral-7B	51.2
Llama-7B	55.1

Table 6.6: AUC detection score on the BOOKSMIA dataset.

We observe that the AUC detection score is higher for larger models, *i.e.*, it is easier to detect if a book was in the pretraining data of a larger LLM. The definition of the MIN-K% PROB measure also means that larger LLMs tend to produce text that is more similar to their pretraining data, such as fiction books, which could help explain their better ASE ratings.

■ Takeaways

Our computed contamination rates lead us to believe that the LLMs we used for our ASE experiments (Section 5.2) were not affected by model contamination. The better performance of larger LLMs for ASG may be partially explained by their tendency to generate text that is more similar to their pretraining data, *e.g.* existing novels.

6.5 Conclusion

6.5.1 Takeaways

1. **LLMs understand the ASE task only partially (Section 6.3):** while they provide explanations that are specific to the evaluated criteria (Section 6.3.1, Section 6.3.2), they struggle to explain their answers with substantiated claims (Section 6.3.3);
2. **Pretraining data helps explain LLM performance at ASG (Section 6.4):** the higher ratings of larger LLMs may be due to their ability to produce output similar to existing books.

6.5.2 Limitations and Future Directions

Since prompt engineering plays a major role on LLM performance [Chen et al., 2023; Liu et al., 2023; Tonmoy et al., 2024], we encourage further testing using more recent prompting techniques such as Chain-of-Thought [Wei et al., 2022b], Automatic Chain-of-Thought [Zhang et al., 2023] or Tree-of-Thought [Yao et al., 2023].

Further discussion on the future of Automatic Story Generation and Automatic Story Evaluation, as well as the future of LLMs, is to be found in Chapter 7.

Chapter 7

Conclusion

*“ Farewell, until our souls cross
paths once more in the boundless
sphere of fate. ”*

Morgana (*The House in Fata
Morgana*)

Contents

7.1 Contributions	134
7.2 Perspectives	135
7.2.1 Designing Specific Automatic Story Evaluation Measures	135
7.2.2 Investigating LLM Performance	136
7.2.3 Assessing and Mitigating the Adverse Societal Impacts of LLMs	137

Chapter 7 Abstract

In this chapter, we summarize our main contributions and findings (Section 7.1), and outline three main research perspectives (Section 7.2): designing specific Automatic Story Evaluation measures, further investigating LLM performance at ASG and ASE, and assessing and mitigating the impact of LLMs on society.

This thesis focused on the specific tasks of Automatic Story Generation and Automatic Story Evaluation. During the three years I spent studying those research fields, Natural Language Processing as a whole has been evolving at a blistering rate, mainly due to the emergence of Large Language Models. Their impressive performance over a large variety of tasks has caused many research directions to become essentially irrelevant (such as the design of task-specific models), but has also brought up new, interesting ones (namely, the questions on model explainability). I found navigating this ever-changing scene both intellectually challenging and stimulating.

In Section 7.1, we summarize our main contributions and findings and, in Section 7.2, we lay out a few promising research directions.

7.1 Contributions

At the beginning of this thesis, we looked at the field of Automatic Story Generation and quickly observed that the problem of evaluating story generation in a satisfactory manner was still very much open.

To tackle this issue, we set out to design an original methodology specifically tailored for story evaluation. We therefore delved into the social sciences literature and designed a comprehensive set of 6 orthogonal human criteria (Relevance, Coherence, Empathy, Surprise, Engagement, and Complexity) for the evaluation of ASG, which are based on the social sciences literature (Section 3.2.3). We also proposed a two-dimensional classification of existing automatic measures so as to identify potential differences in behavior (Section 3.3.1), and we introduced a new meta-evaluation strategy, the overall correlation, which is more adapted to story evaluation than the segment-level correlation (Section 3.4.2).

A major contribution of ours is HANNA, a benchmark of 1,056 stories, both written by humans and generated by 3 ASG-specific systems or 7 pretrained language models, annotated by human raters on six different criteria, and evaluated with 72 automatic measures (Chapter 4). Most notably, we observed that large pre-trained language models produce the best results for ASG (Section 4.3.4). We later augmented HANNA with 480 new stories generated by 4 LLMs, as well as the ~150,000 ratings and explanations from 4 different LLMs asked to evaluate all 1,536 HANNA stories on our human criteria (Section 5.2). We found that LLM stories had at least equal ratings to human stories (Section 4.4.3), as graded by Beluga-13B and Mistral-7B.

Using HANNA, we conducted an extensive meta-evaluation of automatic evaluation measures for Automatic Story Generation. We mainly found that that stronger evaluation measures and that stronger evaluation measures, tailored explicitly for specific criteria of ASG, were needed (Section 5.1.1). We investigated the performance of LLMs at ASE and ASG

task. We showed that LLMs correlated better with human ratings than previous automatic evaluation measures (Section 5.2.1) and that providing detailed guidelines did not lead to improved correlations (Section 5.2.2).

Finally, in order to assess the reliability of our analysis of LLM performance, we performed experiments on LLM explainability (Section 6.3), including namely a user study on LLM explanations that yielded 1,500 annotations (Section 6.3.3). We found that LLMs understand the ASE task only partially and especially struggle to explain their answers with substantiated claims, and that pretraining data helps explain LLM performance at ASG (Section 6.4).

7.2 Perspectives

We identify three main avenues of research pertaining to our work:

1. Designing specific Automatic Story Evaluation measures;
2. Further investigating LLM performance at ASG and ASE;
3. Assessing the impact of LLMs on society.

7.2.1 Designing Specific Automatic Story Evaluation Measures

One of our main findings is that automatic evaluation measures (including LLMs) correlate poorly with human judgment for the overall correlation strategy: this underlines the difficulty of automatically assessing the quality of a specific story. We would like to see more research on the design of specific ASE measures, and we suggest a few interesting directions for each of our human criteria:

1. **Relevance.** We believe summarization-based measures, (such as SUPERT, which performed best for Relevance for overall correlations) are promising contenders, since they can directly compare the prompt with the story.
2. **Coherence.** Various linguistic devices enable coherence, such as causal connectives [Sanders et al., 2007]. Wang and Guo [2014] claim that “A text must [1.] be consistent with context in which it is created [and 2.] have cohesion, that is, all parts in a text must be connected by cohesive devices.” Graph-based measures may be relevant to this criterion, since causal connections can be encoded as a graph [Xu et al., 2019].
3. **Empathy.** The intensity of emotions may be expressed through lexical modalities [Argaman, 2010]: type-token ratio, reducers, intensifiers, repetitions, “feel” derivatives, emotional words, first person singular, similes, exclamatives, etc. Kleres [2011] claims that emotion can be expressed at the word- and sentence-level, but also argues that: “Emotions are inextricably interwoven with the meaning dimension of texts to the point where the distinction between cognition and emotion becomes blurry.” We believe that the ORDSCALE model by Stoehr et al. [2023], which is a Bayesian generative model that handles sentiment as a latent concept with ranking-based features, is an idea worth exploring in greater detail. In particular, the imposed ordering of the parameters of

the categorical distribution may help associate its classification decisions with Likert ratings.

4. **Surprise.** Kumar et al. [2023] show that Bayesian surprise (KL divergence between $t - 1$ and t word probability distributions) is a good predictor of human event segmentation. Chieppe et al. [2022] model the well-made surprise as such: first, during the setup, the audience forms a flawed understanding of the story. A sudden reveal prompts the reader to question their knowledge and, finally, an explanation helps cement a newer and truer understanding. They propose a novel framework to model the audience’s beliefs of a narrative world using approximate Bayesian inference over Markov Logic Networks, and show that their model may have meaningful predictive power. We would like to see a further investigation of these results for ASE.
5. **Engagement.** Bermejo-Berros et al. [2022] show that “experiencing suspense and surprise during the narrative progression influences the level of global enjoyment.” Toubia et al. [2021] introduce three measures of semantic progression that use word embeddings: speed, volume and circuitousness. Piper and Toubia [2023] study non-linearity in storytelling, *i.e.*, the use of literary devices such as time jumps and perspective changes. They show that “non-linearity is weakly negatively associated with reader enjoyment and popularity.” They also use the three aforementioned measures and show that readers prefer stories that cover longer sequences of events with many narrative steps and where newly-acquired information is mainly related to prior information, *i.e.*, with fewer bits of trivia unrelated to the main story points. Those findings may help design a more integrated model of engagement.
6. **Complexity.** Kemper et al. [1990] design four analyses of complexity: structural (interrelationships between story episodes), syntactic (main clauses and subordinates), propositional (density of propositions) and cohesion (references, ellipses, lexical repetitions, substitutions and conjunctions). Petersen et al. [2008] introduce the Index of Narrative Complexity (INC) which “includes categories for rating the complexity of characters, setting, initiating events, internal responses, plans, action/attempts, complications, consequences, narrator evaluations, formulaic markers, temporal markers, and causal adverbial clauses.” We would like to see an attempt at formalizing these features for ASE.

More generally, we did not explore performing supervised training on Large Language Models w.r.t. a specific human criterion (*e.g.* Empathy) for the task of ASE: this option should also be considered.

7.2.2 Investigating LLM Performance

Our latest work is part of the ongoing research on the general ability of LLMs for understanding and thinking. Mahowald et al. [2024] distinguish formal (the statistical features of language) and functional linguistic competence (the ability to use language in the world) and show that LLMs are very successful on formal linguistic tasks but struggle at functional linguistic

tasks. [Bubeck et al. \[2023\]](#) argue that LLMs do display impressive performance at a wide variety of tasks but lack “slow thinking” capabilities, referring to the System 1–System 2 dichotomy introduced by [Kahneman \[2011\]](#). Thus, the high performance of LLMs at ASE should be interpreted with caution: we hypothesize that the “rating” part of our story evaluation experiments could be linked to formal linguistic competence and the fast, automatic System 1, while the “explanation” part would correspond to functional linguistic competence and the slow, conscious System 2. This analogy would explain the good correlations of LLM ratings with human ratings ([Section 5.2.1](#)): the internal criterion of LLMs for story evaluation may be formal quality (vocabulary, syntax, grammar), regardless of the criterion mentioned in the Eval-Prompt. Indeed, our six criteria are mostly orthogonal but not completely independent ([Section 3.2.3](#)): their correlation with one another may be related to the general “System 1” tendency of human raters to favour stories that display better formal qualities. In that sense, LLMs may reflect a human bias towards easy, intuitive thinking. By contrast, the less convincing performance of LLMs at explaining their ratings ([Section 6.3](#)) may highlight their weaker System 2 capabilities.

Other hypotheses have been explored to explain LLM performance. [Schaeffer et al. \[2023\]](#) study whether LLMs do possess *emergent abilities*, *i.e.*, abilities not present in smaller-scale models that are present in larger-scale models. Through a simple mathematical model, they show that such emergent abilities can be an artefact induced by the researcher’s choosing a specific evaluation measure or by the scarcity of test data. [Peeperkorn et al. \[2024\]](#) look at the effect of the temperature parameter on the creativity of generated stories. They perform an empirical analysis and a creativity evaluation with human participants w.r.t. four criteria: novelty, typicality, cohesion, and coherence. They find that temperature has limited influence on LLM output, observing only a weak positive correlation with novelty and a negative correlation with coherence. [Guo et al. \[2024\]](#) discuss the consequences of training language models on synthetic data that was generated by other models, which is becoming an increasingly common practice. They develop a new set of automatic measures that evaluate linguistic diversity and observe a consistent decrease of output diversity, especially through successive iterations.

We believe that more thorough investigations of the inner workings of Large Language Models are needed in order to better identify their potential limitations and areas of improvement.

7.2.3 Assessing and Mitigating the Adverse Societal Impacts of LLMs

There has been an number of researchers raising concerns about the ethical and social risks posed by LLM. [Weidinger et al. \[2022\]](#) identify six areas:

1. Discrimination, Hate speech and Exclusion;
2. Information Hazards;
3. Misinformation Harms;
4. Malicious Uses;
5. Human-Computer Interaction Harms;
6. Environmental and Socioeconomic harms.

For Automatic Story Generation and Evaluation, we believe that **Discrimination, Hate speech and Exclusion** are a major concern. Now that LLMs are already being used for writing contests [Edilivre, 2023] and are being quickly incorporated into the workplace [Humlum and Vestergaard, 2024], people will become more and more exposed to content that will have been automatically generated. In particular, automatically generated literature may soon become a reality or, at least, we may see a rise in entertainment material that is based on generated storylines. Unfortunately, the reproduction of harmful stereotypes is well-documented in language models [Caliskan et al., 2017], and LLMs are also concerned [Bubeck et al., 2023]: GPT-3 was shown to exhibit bias w.r.t. religion, associating “Muslim” with “terrorist” in 23% of test cases [Abid et al., 2021], and gender bias, presenting fictional female characters as less powerful than male counterparts [Lucy and Bamman, 2021]. Furthermore, Gehman et al. [2020] show that language models can degenerate into toxic text even from seemingly innocuous prompts, and that no current method is failsafe against neural toxic degeneration. Several mitigation methods can be employed: curating training data helps make language models fairer, but Denton et al. [2020] warn about the risks of data-gathering operations, which may reinforce forms of social domination or “predatory inclusion”. Martin Jr et al. [2020] suggest to expand metrics and benchmarks to account for social context using community-based system dynamics. Solaiman and Dennison [2021] propose a Process for Adapting Language Models to Society (PALMS), an iterative process to adjust the behavior of a language model to be sensitive to predefined norms.

Finally, we are concerned about the **Environmental and Socioeconomic harms** associated with LLMs. LLMs may be used to produce creative content that would otherwise take time-consuming human labor: this could make creative or innovative work less profitable. Already, Japanese author Rie Kudan has admitted to using CHATGPT for generating around 5% of “The Tokyo Tower of Sympathy”, the book that earned her the Akutagawa Prize, a prestigious literary award. [Choi and Annio, 2024]. Moreover, similar to “patent-busting” [Rimmer, 2016], LLMs may also violate or bypass copyright or through producing content that can be used in place of works which are protected by copyright [Karamolegkou et al., 2023].

On a broader scale than Automatic Story Generation and Evaluation, large amounts of energy are required to train and operate LLMs [Bender et al., 2021], which leads to correspondingly high carbon emissions when said energy comes from fossil fuels. LLMs also require significant volumes of fresh water for the cooling of their data centers, which negatively affects surrounding ecosystems [Mytton, 2021]. These days, several businesses invest more effort in running deep neural network models than in training them: Nvidia stated that 80–90% of their machine learning (ML) workload is reserved for inference, while Amazon Web Services asserted that inference constituted 90% of cloud ML demand. [Patterson et al., 2021]. Approaches to mitigating risks include segmenting LLMs into smaller language models that search and retrieve information from a distinct data corpus [Borgeaud et al., 2022] and targeting efficiency gains during training and inference [Li et al., 2021], but the aggregate effects of reducing energy cost could lead to a Jevons paradox [Jevons, 1866]: more efficient training may foster greater usage of LLMs, resulting in an increase in energy consumption. We believe that those issues can, and should, be discussed at the broader organisational level, *e.g.* as companies shift towards using sustainable energy; and at the public policy level, *e.g.* by

developing more effective carbon pricing and informing the general public about the heavy environmental impact incurred by LLM usage.

Chapter 8

Bibliography

- Abubakar Abid, Maheen Farooqi, and James Zou. 2021. [Persistent anti-Muslim bias in large language models](#). In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery. [138](#)
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. [GPT-4 technical report](#). *ArXiv Preprint*, abs/2303.08774v6. [31](#), [119](#)
- Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. [STORIUM: A dataset and evaluation platform for machine-in-the-loop story generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics. [39](#), [40](#), [41](#), [44](#), [45](#), [48](#), [68](#)
- Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. [Automatic story generation: Challenges and attempts](#). In *Proceedings of the Third Workshop on Narrative Understanding*, pages 72–83, Virtual. Association for Computational Linguistics. [3](#), [38](#)
- Arwa I. Alhussain and Aqil M. Azmi. 2021. [Automatic story generation: A survey of approaches](#). *ACM Computing Surveys*, 54(5). [xi](#), [10](#)
- Sam Altman. 2023. [Planning for AGI and beyond](#). OpenAI Blog. Accessed: 2024-06-15. [31](#)
- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. [Rethinking the agreement in human evaluation tasks](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics. [33](#)
- Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2020. [Story realization: Expanding plot events into sentences](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7375–7382, New York, NY, USA. AAAI Press. [71](#)

- Valentin Amrhein, Sander Greenland, and Blake McShane. 2019. **Scientists rise up against statistical significance**. *Nature*, 567(7748):305–307. [61](#)
- Osnat Argaman. 2010. **Linguistic markers and emotional intensity**. *Journal of Psycholinguistic Research*, 39:89–99. [135](#)
- Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Re. 2023. **Ask me anything: A simple strategy for prompting language models**. In *The Eleventh International Conference on Learning Representations*. [86](#)
- Ron Artstein and Massimo Poesio. 2008. **Inter-coder agreement for computational linguistics**. *Computational Linguistics*, 34(4):555–596. [80](#)
- Byung-Chull Bae, Suji Jang, Youngjune Kim, and Seyoung Park. 2021. **A preliminary survey on story interestingness: Focusing on cognitive and emotional interest**. In *International Conference on Interactive Digital Storytelling*, pages 447–453. Springer. [47](#), [48](#), [49](#)
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. **Neural machine translation by jointly learning to align and translate**. In *3rd International Conference on Learning Representations (ICLR) 2015*, San Diego, CA, USA. [25](#), [26](#)
- He Bai, Peng Shi, Jimmy Lin, Luchen Tan, Kun Xiong, Wen Gao, Jie Liu, and Ming Li. 2021. **Semantics of the unwritten: The effect of end of paragraph and sequence tokens on text generation with GPT2**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 148–162, Online. Association for Computational Linguistics. [40](#), [41](#), [42](#), [44](#), [45](#), [48](#), [73](#), [74](#), [75](#), [76](#), [82](#)
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics. [50](#), [51](#)
- Randolph T. Barker and Kim Gower. 2010. **Strategic application of storytelling in organizations: Toward effective communication in a diverse world**. *The Journal of Business Communication* (1973), 47(3):295–312. [2](#)
- Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James R. Glass. 2019. **Identifying and controlling important neurons in neural machine translation**. In *7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA. OpenReview.net. [123](#)
- Morteza Behrooz, Justus Robertson, and Arnav Jhala. 2019. **Story quality as a matter of perception: Using word embeddings to estimate cognitive interest**. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2019, October 8-12*, volume 15, pages 3–9, Atlanta, Georgia, USA. AAAI Press. [49](#)

- Anja Belz and Eric Kow. 2011. **Discrete vs. continuous rating scales for language evaluation in NLP**. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–235, Portland, Oregon, USA. Association for Computational Linguistics. 32
- Anja Belz and Ehud Reiter. 2006. **Comparing automatic and human evaluation of NLG systems**. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics. 32
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. **On the dangers of stochastic parrots: Can language models be too big?** In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 610–623, New York, NY, USA. Association for Computing Machinery. 138
- Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. **A neural probabilistic language model**. In *Advances in Neural Information Processing Systems*, volume 13, pages 932–938, Denver, CO, USA. MIT Press. 23
- Yoav Benjamini and Yosef Hochberg. 1995. **Controlling the false discovery rate: A practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300. 61
- Jesús Bermejo-Berros, Jaime Lopez-Diez, and Miguel Angel Gil Martínez. 2022. **Inducing narrative tension in the viewer through suspense, surprise, and curiosity**. *Poetics*, 93:101664. 136
- Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. **Re-evaluating evaluation in text summarization**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics. 34, 59, 102
- John B. Black and Gordon H. Bower. 1980. **Story understanding as problem-solving**. *Poetics*, 9(1):223–250. 12
- John B. Black and Robert Wilensky. 1979. **An evaluation of story grammars**. *Cognitive Science*, 3(3):213–229. 12
- Margaret A. Boden. 2009. **Computer models of creativity**. *AI Magazine*, 30(3):23. 2
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. **Enriching word vectors with subword information**. *Transactions of the Association for Computational Linguistics*, 5:135–146. 23
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving,

- Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. **Improving language models by retrieving from trillions of tokens**. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240, Baltimore, Maryland, USA. PMLR. 138
- Barbaros Bostan and Tim Marsh. 2010. **The ‘interactive’ of interactive storytelling: Customizing the gaming experience**. In *Entertainment Computing-ICEC 2010: 9th International Conference, ICEC 2010, Seoul, Korea, September 8-11, 2010. Proceedings 9*, pages 472–475, Berlin, Heidelberg. Springer. 2
- Faeze Brahman and Snigdha Chaturvedi. 2020. **Modeling protagonist emotions for emotion-aware storytelling**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics. 41, 42, 45, 72
- Faeze Brahman, Alexandru Petrusca, and Snigdha Chaturvedi. 2020. **Cue me in: Content-inducing approaches to interactive story generation**. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 588–597, Suzhou, China. Association for Computational Linguistics. 45, 74
- William F. Brewer and Edward H. Lichtenstein. 1980. **Event schemas, story schemas, and story grammars**. *Center for the Study of Reading Technical Report*, 197. 11
- Selmer Bringsjord and David Ferrucci. 1999. *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*. Psychology Press. 11
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. **Language models are few-shot learners**. In *Advances in Neural Information Processing Systems*, volume 33, Online. 3, 31, 82, 86, 91
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. **Sparks of artificial general intelligence: Early experiments with GPT-4**. *ArXiv Preprint*, abs/2303.12712v5. 31, 137, 138
- Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. **Semantics derived automatically from language corpora contain human-like biases**. *Science*, 356(6334):183–186. 138
- Chris Callison-Burch. 2009. **Fast, cheap, and creative: Evaluating translation quality using Amazon’s Mechanical Turk**. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics. 32

- Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. [Re-evaluating the role of Bleu in machine translation research](#). In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics. [34](#)
- Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. [YAKE! Keyword extraction from single documents using multiple local features](#). *Information Sciences*, 509:257–289. [125](#)
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. [Evaluation of text generation: A survey](#). *ArXiv Preprint*, abs/2006.14799v2. [82](#)
- Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. [The price of debiasing automatic metrics in natural language evaluation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics. [33](#)
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. [Art or artifice? Large language models and the false promise of creativity](#). In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA. Association for Computing Machinery. [85](#)
- Ronan Champagnat, Guylain Delmas, and Michel Augeraud. 2010. [A storytelling model for educational games: Hero’s interactive journey](#). *International Journal of Technology Enhanced Learning*, 2(1-2):4–20. [11](#)
- Shweta Chauhan and Philemon Daniel. 2023. [A comprehensive survey on various fully automatic machine translation evaluation metrics](#). *Neural Processing Letters*, 55(9):12663–12717. [51](#)
- Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. [Evaluating question answering evaluation](#). In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics. [34](#)
- Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. [Unleashing the potential of prompt engineering in large language models: A comprehensive review](#). *ArXiv Preprint*, abs/2310.14735v4. [131](#)
- Stanley F. Chen, Douglas Beeferman, and Roni Rosenfeld. 1998. [Evaluation metrics for language models](#). [15](#)
- Stanley F. Chen and Ronald Rosenfeld. 2000. [A survey of smoothing techniques for ME models](#). *IEEE transactions on Speech and Audio Processing*, 8(1):37–50. [16](#)
- Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. [Of human criteria and automatic metrics: A benchmark of the evaluation of story generation](#). In *Proceedings of*

- the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. 7, 60
- Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. 2024. **Do language models enjoy their own stories? Prompting large language models for automatic story evaluation**. *Transactions of the Association for Computational Linguistics*, 12:1122–1142. 7, 60
- Patrick Chieppe, Penny Sweetser, and Eryn Newman. 2022. **Bayesian modelling of the well-made surprise**. In *Proceedings of the 13th International Conference on Computational Creativity*, pages 126–135, Bozen-Bolzano, Italy. 136
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. **Learning phrase representations using RNN encoder–decoder for statistical machine translation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics. 24, 25
- Christy Choi and Francesca Annio. 2024. **The winner of a prestigious Japanese literary award has confirmed AI helped write her book**. *CNN*. 138
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. **PaLM: Scaling language modeling with pathways**. *Journal of Machine Learning Research*, 24(240):1–113. 31, 82
- Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. **All that’s ‘human’ is not gold: Evaluating human evaluation of generated text**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics. 82
- Elizabeth Clark, Yangfeng Ji, and Noah A. Smith. 2018. **Neural text generation in stories using entity representations as context**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260, New Orleans, Louisiana. Association for Computational Linguistics. 40, 44, 70

- Elizabeth Clark and Noah A. Smith. 2021. **Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3566–3575, Online. Association for Computational Linguistics. [40](#), [44](#)
- Jacob Cohen. 1960. **A coefficient of agreement for nominal scales**. *Educational and Psychological Measurement*, 20(1):37–46. [33](#), [112](#)
- Benjamin N. Colby. 1973. **A partial grammar of Eskimo folktales**. *American Anthropologist*, 75(3):645–662. [11](#)
- Pierre Colombo, Chloé Clavel, and Pablo Piantanida. 2022a. **InfoLM: A new metric to evaluate summarization & data2text generation**. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 10554–10562, Online. AAAI Press. [50](#), [55](#)
- Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Cléménçon. 2022b. **What are the best systems? New perspectives on NLP benchmarking**. In *Advances in Neural Information Processing Systems*, volume 35, New Orleans, LA, USA. [110](#)
- Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021. **Automatic text evaluation through the lens of Wasserstein barycenters**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. [34](#), [50](#), [54](#)
- Richard Craggs and Mary McGee Wood. 2005. **Evaluating discourse and dialogue coding schemes**. *Computational Linguistics*, 31(3):289–296. [80](#)
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. **The PASCAL recognising textual entailment challenge**. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190, Berlin, Heidelberg. Springer. [64](#)
- Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. **Language modeling with gated convolutional networks**. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 933–941, Sydney, NSW, Australia. PMLR. [70](#)
- James Deese. 1978. **Thought into speech: Linguistic rules and psychological limitations in processing information determine how we put our ideas into words**. *American Scientist*, 66(3):314–321. [46](#)
- Natalie Dehn. 1981. **Story generation after TALE-SPIN**. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI’81*, page 16–18, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. [12](#)

- John E. Dennis Jr. and Robert B. Schnabel. 1996. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics. 22
- Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. [Bringing the people back in: Contesting benchmark machine learning datasets](#). *ArXiv Preprint*, abs/2007.07399v1. 138
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. 3, 30, 50, 53, 57, 123
- Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. [Handling divergent reference texts when evaluating table-to-text generation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics. 33, 34
- Robert Dickman. 2003. [The four elements of every successful story](#). *Reflections - Society for Organizational Learning*, 4(3):51–58. 46, 48, 49
- Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. [Is GPT-3 a good data annotator?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics. 85
- George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. [The automatic content extraction \(ACE\) program – tasks, data, and evaluation](#). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal. European Language Resources Association (ELRA). 64
- Finale Doshi-Velez and Been Kim. 2017. [Towards a rigorous science of interpretable machine learning](#). *ArXiv Preprint*, abs/1702.08608v2. 122
- Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. [Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics. 126
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics. 103

- John C. Duchi, Elad Hazan, and Yoram Singer. 2010. [Adaptive subgradient methods for online learning and stochastic optimization](#). In *COLT 2010 - The 23rd Conference on Learning Theory*, pages 257–269, Haifa, Israel. Omnipress. 22
- Olive Jean Dunn. 1959. [Estimation of the medians for dependent variables](#). *The Annals of Mathematical Statistics*, 30(1):192–197. 61
- Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. [Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge](#). *Computer Speech & Language*, 59:123–156. 32
- Edilivre. 2023. [Concours de nouvelles 2023](#). Edilivre. Accessed: 2023-10-12. 82, 138
- Desmond Elliott and Frank Keller. 2014. [Comparing automatic evaluation measures for image description](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland. Association for Computational Linguistics. 34
- Majigsuren Enkhsaikhan, Wei Liu, Eun-Jung Holden, and Paul Duuring. 2021. [Auto-labelling entities in low-resource text: A geological case study](#). *Knowledge and Information Systems*, 63:695–715. 85
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating summarization evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409. 34, 50, 56, 96
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics. xv, 40, 44, 48, 58, 67, 70, 71, 74, 75, 130
- Angela Fan, Mike Lewis, and Yann Dauphin. 2019. [Strategies for structuring story generation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics. 41, 45, 71
- Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. [Transformer-based conditional variational autoencoder for controllable story generation](#). *ArXiv Preprint*, abs/2101.00828v2. 73, 74
- Aris Fergadis and Benedikt Scheffler. 2022. [Chance-corrected agreement coefficients](#). 128
- Joseph L. Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76(5):378. 33, 112
- Yang Gao, Wei Zhao, and Steffen Eger. 2020. [SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics. 50, 57

- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. **RealToxicityPrompts: Evaluating neural toxic degeneration in language models**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics. [138](#)
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. **Convolutional sequence to sequence learning**. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, Sydney, NSW, Australia. PMLR. [70](#), [71](#)
- Pablo Gervás, Belén Díaz-Agudo, Federico Peinado, and Raquel Hervás. 2005. **Story plot generation based on CBR**. In *Applications and Innovations in Intelligent Systems XII*, pages 33–46, London. Springer London. [13](#)
- Sarik Ghazarian, Zixi Liu, Akash S M, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2021. **Plot-guided adversarial example construction for evaluating open-domain story generation**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4334–4344, Online. Association for Computational Linguistics. [42](#), [45](#)
- Andrew R. Gilpin. 1993. **Table for conversion of Kendall’s tau to Spearman’s rho within the context of measures of magnitude of effect for meta-analysis**. *Educational and Psychological Measurement*, 53(1):87–92. [60](#)
- Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. **Deep sparse rectifier neural networks**. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323. JMLR Workshop and Conference Proceedings, PMLR. [17](#)
- Yoav Goldberg. 2016. **A primer on neural network models for natural language processing**. *Journal of Artificial Intelligence Research*, 57:345–420. [22](#)
- Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. **Content planning for neural story generation with aristotelian rescoring**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics. [40](#), [42](#), [44](#), [45](#), [48](#), [73](#), [74](#)
- Joshua T. Goodman. 2001. **A bit of progress in language modeling**. *Computer Speech & Language*, 15(4):403–434. [16](#)
- Cyril Goutte and Eric Gaussier. 2005. **A probabilistic interpretation of precision, recall and F-score, with implication for evaluation**. In *Advances in Information Retrieval*, pages 345–359, Berlin, Heidelberg. Springer Berlin Heidelberg. [103](#)
- Yvette Graham and Timothy Baldwin. 2014. **Testing for significance of increased correlation with human judgment**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics. [61](#)

- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. [Can machine translation systems be evaluated by the crowd alone](#). *Natural Language Engineering*, 23(1):3–30. [33](#), [80](#)
- Dieter Grasbon and Norbert Braun. 2001. [A morphological approach to interactive storytelling](#). In *Proceedings of the Conference on Artistic, Cultural and Scientific Aspects of Experimental Media Spaces (CAST) 2001*, pages 337–340. Fraunhofer-Institut für Graphische Datenverarbeitung IGD. [11](#)
- Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. [Neural turing machines](#). *ArXiv Preprint*, abs/:1410.5401v2. [26](#)
- Ralph Grishman and Beth Sundheim. 1996. [Message Understanding Conference- 6: A brief history](#). In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. [64](#)
- Max Grusky, Mor Naaman, and Yoav Artzi. 2018. [Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics. [50](#), [56](#)
- Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. [A knowledge-enhanced pretraining model for commonsense story generation](#). *Transactions of the Association for Computational Linguistics*, 8:93–108. [40](#), [41](#), [44](#), [45](#), [72](#), [74](#)
- Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021a. [Long text generation by modeling sentence-level and discourse-level coherence](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6379–6393, Online. Association for Computational Linguistics. [39](#), [41](#), [44](#), [48](#), [73](#), [74](#), [75](#)
- Jian Guan, Yansen Wang, and Minlie Huang. 2019. [Story ending generation with incremental encoding and commonsense knowledge](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 6473–6480, Honolulu, Hawaii, USA. AAAI Press. [40](#), [41](#), [44](#), [45](#)
- Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021b. [OpenMEVA: A benchmark for evaluating open-ended story generation metrics](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online. Association for Computational Linguistics. [42](#), [45](#), [58](#), [68](#), [69](#)
- Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. [The curious decline of linguistic diversity: Training language models on synthetic text](#). In *Findings of*

- the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics. 137
- Kilem Li Gwet. 2008. **Computing inter-rater reliability and its variance in the presence of high agreement**. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48. 128
- Ivan Habernal and Iryna Gurevych. 2017. **Argumentation mining in user-generated web discourse**. *Computational Linguistics*, 43(1):125–179. 80
- Pierre Hadot. 2002. *What Is Ancient Philosophy?* Harvard University Press. 46
- Kevin A. Hallgren. 2012. **Computing inter-rater reliability for observational data: An overview and tutorial**. *Tutorials in Quantitative Methods for Psychology*, 8(1):23. 80, 112
- Patrik Haslum. 2012. Narrative planning: Compilations to classical planning. *Journal of Artificial Intelligence Research*, 44:383–395. 12
- Martin Hassel. 2004. **Evaluation of automatic text summarization**. *Licentiate Thesis, KTH Royal Institute of Technology*, page 8. 33
- Andrew F. Hayes and Klaus Krippendorff. 2007. **Answering the call for a standard reliability measure for coding data**. *Communication Methods and Measures*, 1(1):77–89. 33, 112
- Donald Olding Hebb. 1949. *The Organization of Behavior: A Neuropsychological Theory*. Wiley. 17
- Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. **Deep read: A reading comprehension system**. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 325–332, College Park, Maryland, USA. Association for Computational Linguistics. 64
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. **Long short-term memory**. *Neural Computation*, 9(8):1735–1780. 20
- John J. Hopfield. 1982. **Neural networks and physical systems with emergent collective computational abilities**. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558. 17
- Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. 2020. **Linguistic appropriateness and pedagogic usefulness of reading comprehension questions**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1753–1762, Marseille, France. European Language Resources Association. 32
- Jeremy Howard and Sebastian Ruder. 2018. **Universal language model fine-tuning for text classification**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics. 72

- Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. **Visual storytelling**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics. [xv](#), [65](#), [66](#)
- Anders Humlum and Emilie Vestergaard. 2024. **The adoption of ChatGPT**. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 50. [138](#)
- Asghar Iran-Nejad. 1987. **Cognitive and affective causes of interest and liking**. *Journal of Educational Psychology*, 79(2):120–130. [49](#)
- Mohieddin Jafari and Naser Ansari-Pour. 2019. **Why, when and how to adjust your P values?** *Cell Journal (Yakhteh)*, 20(4):604–607. [61](#)
- William Stanley Jevons. 1866. *The Coal Question; An Inquiry Concerning the Progress of the Nation and the Probable Exhaustion of our Coal-mines*. Macmillan. [138](#)
- Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2020. **Narrative text generation with a latent discrete plan**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3637–3650, Online. Association for Computational Linguistics. [39](#), [40](#), [44](#), [48](#), [71](#), [74](#)
- Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. 2017. **Dynamic entity representations in neural language models**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839, Copenhagen, Denmark. Association for Computational Linguistics. [70](#)
- Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. **Is BERT really robust? A strong baseline for natural language attack on text classification and entailment**. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8018–8025, New York, NY, USA. AAAI Press. [123](#)
- Kun Jing and Jungang Xu. 2019. **A survey on neural network language models**. *ArXiv Preprint*, abs/1906.03591v2. [23](#)
- Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux. [137](#)
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. **Scaling laws for neural language models**. *ArXiv Preprint*, abs/2001.08361v1. [31](#)
- Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. **Copyright violations and large language models**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore. Association for Computational Linguistics. [138](#)

- Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. [The perils of using Mechanical Turk to evaluate open-ended text generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. [76](#), [80](#)
- Bilal Kartal, John Koenig, and Stephen J. Guy. 2014. [User-driven narrative variation in large story domains using Monte Carlo tree search](#). In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, pages 69–76. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS). [13](#)
- Suzanne Keen. 2007. *Empathy and the Novel*. Oxford University Press on Demand. [49](#)
- John G. Kemeny. 1959. [Mathematics without numbers](#). *Daedalus*, 88(4):577–591. [110](#)
- Susan Kemper, Shannon Rash, Donna Kynette, and Suzanne Norman. 1990. [Telling stories: The structure of adults’ narratives](#). *European Journal of Cognitive Psychology*, 2(3):205–228. [136](#)
- Maurice G. Kendall. 1938. [A new measure of rank correlation](#). *Biometrika*, 30(1/2):81–93. [58](#), [60](#)
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. [CTRL: A conditional transformer language model for controllable generation](#). *ArXiv Preprint*, abs/1909.05858v2. [75](#)
- Ahmed Khalifa, Gabriella A. B. Barros, and Julian Togelius. 2017. [DeepTingle](#). In *Proceedings of the Eighth International Conference on Computational Creativity (ICCC)*, pages 167–174, Atlanta, Georgia, USA. Association for Computational Creativity (ACC). [39](#), [40](#), [44](#), [48](#), [69](#)
- Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. [The \(un\)reliability of saliency methods](#). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. [123](#)
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA. [22](#)
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. [Understanding the effects of RLHF on LLM generalisation and diversity](#). In *The Twelfth International Conference on Learning Representations (ICLR)*, Vienna, Austria. OpenReview.net. [119](#)
- Jochen Kleres. 2011. [Emotions and narrative analysis: A methodological approach](#). *Journal for the Theory of Social Behaviour*, 41(2):182–202. [135](#)
- Reinhard Kneser and Hermann Ney. 1995. [Improved backing-off for M-gram language modeling](#). In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184. IEEE. [16](#)

- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. [The NarrativeQA reading comprehension challenge](#). *Transactions of the Association for Computational Linguistics*, 6:317–328. [64](#)
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics. [103](#)
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc. [86](#)
- Martin Kreiswirth. 2000. [Merely telling stories? Narrative and knowledge in the human sciences](#). *Poetics Today*, 21(2):293–318. [2](#)
- Manoj Kumar, Ariel Goldstein, Sebastian Michelmann, Jeffrey M. Zacks, Uri Hasson, and Kenneth A. Norman. 2023. [Bayesian surprise predicts human event segmentation in story listening](#). *Cognitive Science*, 47(10):e13343. [136](#)
- Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. [From word embeddings to document distances](#). In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966, Lille, France. JMLR.org. [54](#)
- Ben Kybartas and Rafael Bidarra. 2017. [A survey on story generation techniques for authoring computational narratives](#). *IEEE Transactions on Computational Intelligence and AI in Games*, 9(3):239–253. [10](#)
- Ben Kybartas and Clark Verbrugge. 2014. [Analysis of ReGEN as a graph-rewriting system for quest generation](#). *IEEE Transactions on Computational Intelligence and AI in Games*, 6(2):228–242. [11](#)
- William Labov and Joshua Waletzky. 1997. [Narrative analysis: Oral versions of personal experience](#). *Journal of Narrative & Life History*, 7(1-4):3–38. [46](#)
- George Lakoff. 1972. [Structural complexity in fairy tales](#). *UC Berkeley*. [11](#)
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174. [80](#)
- Yann Le Cun, Lawrence D. Jackel, Brian Boser, John S. Denker, Hans Peter Graf, Isabelle Guyon, Don Henderson, Richard E. Howard, and William Hubbard. 1990. [Handwritten digit recognition: Applications of neural network chips and automatic learning](#). In *Neurocomputing*, volume 68 of *NATO ASI Series*, pages 303–318, Berlin, Heidelberg. Springer. [17](#)
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. [Best practices for the human evaluation of automatically generated text](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics. [39](#)

- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. [The Winograd schema challenge](#). In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561, Rome, Italy. AAAI Press. 64
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. 31, 57, 73
- Zhuohan Li, Siyuan Zhuang, Shiyuan Guo, Danyang Zhuo, Hao Zhang, Dawn Song, and Ion Stoica. 2021. [TeraPipe: Token-level pipeline parallelism for training large-scale language models](#). In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 6543–6552, Online. PMLR. 138
- Zongxia Li, Paiheng Xu, Fuxiao Liu, and Hyemi Song. 2023. [Towards understanding in-context learning with contrastive demonstrations and saliency maps](#). *ArXiv Preprint*, abs/2307.05052v4. 123
- Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. [OpenOrca: An open dataset of GPT augmented FLAN reasoning traces](#). 91
- Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. [Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators](#). In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23, New York, NY, USA. Association for Computing Machinery. 119
- Rensis Likert. 1932. [A technique for the measurement of attitudes](#). *Archives of Psychology*. 32
- Worasa Limpanadusadee, Proadpran Punyabukkana, Atiwong Suchato, and Onintra Poo-brasert. 2014. [Text corpus for natural language story-telling sentence generation: A design and evaluation](#). In *Proceedings of the 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 80–85. 16
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. 49, 50, 51
- Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. [Open Sesame: Getting inside BERT's linguistic knowledge](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics. 123
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation](#). In *Proceedings of the 2016 Conference*

- on *Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics. [34](#)
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#). *ACM Computing Surveys*, 55(9). [131](#)
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *ArXiv Preprint*, abs/1907.11692v1. [30](#), [75](#)
- Annie Louis and Charles Sutton. 2018. [Deep Dungeons and Dragons: Learning character-action interactions from role-playing game transcripts](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713, New Orleans, Louisiana. Association for Computational Linguistics. [xv](#), [67](#), [68](#)
- Li Lucy and David Bamman. 2021. [Gender and representation bias in GPT-3 generated stories](#). In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics. [138](#)
- Anna Lundqvist, Veronica Liljander, Johanna Gummerus, and Allard van Riel. 2013. [The impact of storytelling on the consumer brand experience: The case of a firm-originated story](#). *Journal of Brand Management*, 20(4):283–297. [2](#)
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. [Effective approaches to attention-based neural machine translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics. [26](#)
- Stian Lydersen. 2024. [Adjustment of p values for multiple hypotheses: why, when and how](#). *Annals of the Rheumatic Diseases*. [61](#)
- Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. [Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics. [34](#), [59](#)
- Isabel Machado, Ana Paiva, and Paul Brna. 2001. [Real characters in virtual stories](#). In *Virtual Storytelling Using Virtual Reality Technologies for Storytelling*, pages 127–134, Berlin, Heidelberg. Springer. [11](#)
- Inbal Magar and Roy Schwartz. 2022. [Data contamination: From memorization to exploitation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics. [129](#)

- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. [Dissociating language and thought in large language models](#). *Trends in Cognitive Sciences*. 136
- Jean M. Mandler and Nancy S. Johnson. 1977. [Remembrance of things parsed: Story structure and recall](#). *Cognitive Psychology*, 9(1):111–151. 11
- Huanru Henry Mao, Bodhisattwa Prasad Majumder, Julian McAuley, and Garrison Cottrell. 2019. [Improving neural story generation by targeted common sense grounding](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5988–5993, Hong Kong, China. Association for Computational Linguistics. 72, 74
- Pierre Maranda. 1985. Semiography and artificial intelligence. *International Semiotic Spectrum*, 4:1–3. 11
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a large annotated corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330. 64
- Lara J. Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O. Riedl. 2018. [Event representations for automated story generation with deep neural nets](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 868–875, New Orleans, Louisiana, USA. AAAI Press. 69
- Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. 2020. [Participatory problem formulation for fairer machine learning through community based system dynamics](#). *ArXiv Preprint*, abs/2005.07572v3. 138
- Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. [Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics. 34, 80
- Allyssa McCabe and Carole Peterson. 1984. [What makes a good story](#). *Journal of Psycholinguistic Research*, 13(6):457–480. 46, 48, 49
- John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 2006. [A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955](#). *AI Magazine*, 27(4):12. 2
- Warren S. McCulloch and Walter Pitts. 1943. [A logical calculus of the ideas immanent in nervous activity](#). *The Bulletin of Mathematical Biophysics*, 5(4):115–133. 17
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. [UMAP: Uniform manifold approximation and projection](#). *Journal of Open Source Software*, 3(29):861. 124

- Neil McIntyre and Mirella Lapata. 2010. **Plot induction and evolutionary search for story generation**. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1562–1572, Uppsala, Sweden. Association for Computational Linguistics. [xi](#), [13](#), [14](#)
- Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. **Sources of hallucination by large language models on inference tasks**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics. [123](#)
- Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. 2019. **Abandon statistical significance**. *The American Statistician*, 73(sup1):235–245. [61](#)
- James R. Meehan. 1977. **TALE-SPIN, an interactive program that writes stories**. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI’77*, pages 91–98, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. [12](#)
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. **Efficient estimation of word representations in vector space**. *ArXiv Preprint*, abs/1301.3781v3. [xi](#), [23](#), [24](#), [50](#)
- Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. **Recurrent neural network based language model**. In *Interspeech*, volume 2 issue 3, pages 1045–1048. Makuhari. [23](#)
- Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. **Distributed representations of words and phrases and their compositionality**. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119, Lake Tahoe, Nevada, United States. [23](#), [33](#), [50](#)
- George A. Miller. 1992. **WordNet: A lexical database for English**. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. [33](#)
- Marvin Minsky and Seymour A. Papert. 1969. *An Introduction to Computational Geometry*. MIT Press. [17](#)
- Jihyung Moon. 2019. **Significance test of increase in correlation for NLP evaluations in Python**. [61](#)
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. **A corpus and cloze evaluation for deeper understanding of commonsense stories**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics. [xv](#), [58](#), [65](#), [66](#), [71](#)
- Stefanie Muff, Erlend B. Nilsen, Robert B. O’Hara, and Chloé R. Nater. 2022. **Rewriting results sections in the language of evidence**. *Trends in Ecology & Evolution*, 37(3):203–210. [xv](#), [61](#), [62](#)

- Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. **Orca: Progressive learning from complex explanation traces of GPT-4**. *ArXiv Preprint*, abs/2306.02707v1. 91
- Roger B. Myerson. 2013. **Fundamentals of social choice theory**. *Quarterly Journal of Political Science*, 8(3):305–337. 110
- David Mytton. 2021. **Data centre water consumption**. *npj Clean Water*, 4(1):11. 138
- Vinod Nair and Geoffrey E. Hinton. 2010. **Rectified linear units improve restricted boltzmann machines**. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, June 21–24, 2010, Haifa, Israel, pages 807–814. Omnipress. 17
- Preksha Nema and Mitesh M. Khapra. 2018. **Towards a better metric for evaluating question generation systems**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics. 34
- Izaak Neutelings. 2021. **Neural networks**. xi, 19
- Roger Newson. 2002. **Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences**. *The Stata Journal*, 2(1):45–64. 60
- Jun-Ping Ng and Viktoria Abrecht. 2015. **Better summarization evaluation with word embeddings for ROUGE**. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics. 34, 50, 53
- Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. **Why we need new evaluation metrics for NLG**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics. 33, 34, 96
- TeongJoo Ong and John J. Leggett. 2004. **A genetic algorithm approach to interactive narrative generation**. In *Proceedings of the Fifteenth ACM Conference on Hypertext and Hypermedia*, HYPERTEXT ’04, pages 181—182, New York, NY, USA. Association for Computing Machinery. 13
- Curtis Onuczko, Duane Szafron, and Jonathan Schaeffer. 2008. Stop getting side-tracked by side-quests. *AI Game Programming Wisdom*, 4:513–528. 11
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. **Training language models to follow instructions with human feedback**. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc. 31, 91

- Paolo Paolini and Nicoletta Di Blas. 2014. **Storytelling for cultural heritage**. In *Innovative Technologies in Urban Mapping: Built Space and Mental Space*, pages 33–45. Springer International Publishing. [2](#)
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **BLEU: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. [4](#), [33](#), [49](#), [50](#), [51](#)
- Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. **A plug-and-play method for controlled text generation**. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics. [41](#), [43](#), [44](#), [45](#), [74](#)
- David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. **Carbon emissions and large neural network training**. *ArXiv Preprint*, abs/2104.10350v3. [138](#)
- Karl Pearson. 1895. **VII. Note on regression and inheritance in the case of two parents**. *Proceedings of the Royal Society of London*, 58(347-352):240–242. [60](#)
- Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. **Is temperature the creativity parameter of large language models?** *ArXiv preprint*, abs/2405.00492v1. [137](#)
- Lyn Pemberton. 1989. **A modular approach to story generation**. In *Fourth Conference of the European Chapter of the Association for Computational Linguistics*, Manchester, England. Association for Computational Linguistics. [11](#)
- Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. **Towards controllable story generation**. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics. [40](#), [44](#), [70](#), [74](#)
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. **GloVe: Global vectors for word representation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. [23](#)
- Rafael Pérez y Pérez and Mike Sharples. 2001. **MEXICA: A computer model of a cognitive account of creative writing**. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(2):119–139. [13](#)
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. **Deep contextualized word representations**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. [26](#)

- Douglas B. Petersen, Sandra Laing Gillam, and Ronald B. Gillam. 2008. [Emerging procedures in narrative assessment: The index of narrative complexity](#). *Topics in Language Disorders*, 28(2):115–130. 136
- Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. [Learning to score system summaries for better content selection evaluation](#). In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics. 50, 55
- Jason Phang, Thibault F  vry, and Samuel R. Bowman. 2018. [Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks](#). *ArXiv Preprint*, abs/1811.01088v2. 72
- Andrew Piper and Olivier Toubia. 2023. [A quantitative study of non-linearity in storytelling](#). *Poetics*, 98:101793. 136
- Maja Popovi  . 2015. [ChrF: Character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics. 50, 52
- Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas Press. 4, 10, 11
- Muhammad Aasim Qureshi, Muhammad Asif, Mohd Fadzil Hassan, Ghulam Mustafa, Muhammad Khurram Ehsan, Aasim Ali, and Unaza Sajid. 2022. [A novel auto-annotation technique for aspect level sentiment analysis](#). *Computers, Materials and Continua*, 70(3):4987–5004. 85
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#). 29, 72, 75
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. [Language models are unsupervised multitask learners](#). *OpenAI Blog*, 1(8):9. 29, 72, 73, 75, 86
- Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. [Compressive transformers for long-range sequence modelling](#). In *8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia. OpenReview.net. 68
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67. 31
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics. 64
- William Lowell Randall. 1999. [Narrative intelligence and the novelty of our lives](#). *Journal of Aging Studies*, 13(1):11–28. 49

- Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. **PlotMachines: Outline-conditioned generation with dynamic plot state tracking**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics. [42](#), [45](#), [72](#), [74](#)
- Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. **Learning from crowds**. *Journal of Machine Learning Research*, 11(43):1297–1322. [33](#)
- Nils Reimers and Iryna Gurevych. 2019. **Sentence-BERT: Sentence embeddings using Siamese BERT-networks**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. [124](#)
- Ehud Reiter. 2018. **A structured review of the validity of BLEU**. *Computational Linguistics*, 44(3):393–401. [34](#)
- Laria Reynolds and Kyle McDonell. 2021. **Prompt programming for large language models: Beyond the few-shot paradigm**. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7. [86](#)
- Mark O. Riedl and Robert M. Young. 2004. **An intent-driven planner for multi-agent story generation**. In *International Joint Conference on Autonomous Agents and Multiagent Systems*, volume 2, pages 186–193. IEEE Computer Society. [12](#)
- Mark O. Riedl and Robert M. Young. 2010. **Narrative planning: Balancing plot and character**. *Journal of Artificial Intelligence Research*, 39:217–268. [12](#)
- Matthew Rimmer. 2016. **Patent-busting: The public patent foundation, gene patents and the seed wars**. In *The Intellectual Property and Food Project*. Routledge. [138](#)
- Alan Ritter, Colin Cherry, and William B. Dolan. 2011. **Data-driven response generation in social media**. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics. [80](#)
- Hanna-Riikka Roine. 2016. *Imaginative, Immersive and Interactive Engagements. The Rhetoric of Worldbuilding in Contemporary Speculative Fiction*. Number 1695 in Acta Electronica Universitatis Tampereensis. Tampere University Press. [49](#)
- Frank Rosenblatt. 1958. **The perceptron: A probabilistic model for information storage and organization in the brain**. *Psychological Review*, 65(6):386–408. [17](#)
- Ronald Rosenfeld. 2000. **Two decades of statistical language modeling: Where do we go from here?** *Proceedings of the IEEE*, 88(8):1270–1278. [14](#), [16](#)

- Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. **Leveraging pre-trained checkpoints for sequence generation tasks**. *Transactions of the Association for Computational Linguistics*, 8:264–280. 75
- David E. Rumelhart. 1975. **Notes on a schema for stories**. In *Representation and Understanding*, pages 211–236. Morgan Kaufmann, San Diego. 11, 46
- David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1988. **Learning internal representations by error propagation**. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, volume 1, page 0. MIT Press. 17, 22
- James Ryan. 2017. **Grimes’ fairy tales: A 1960s story generator**. In *Interactive Storytelling: 10th International Conference on Interactive Digital Storytelling (ICIDS)*, pages 89–103, Funchal, Madeira, Portugal. Springer International Publishing. xv, 4, 10, 11
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. **A systematic survey of prompt engineering in large language models: Techniques and applications**. *ArXiv Preprint*, abs/2402.07927v1. 86
- Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. **A survey of evaluation metrics used for NLG systems**. *ACM Computing Surveys (CSUR)*, 55(2):1–39. 32, 51
- Geoffrey Sampson and Babarczy Anna. 2008. **Definitional and human constraints on structural annotation of english**. *Natural Language Engineering*, 14(4):471–494. 33
- Arthur L. Samuel. 1959. Machine learning. *The Technology Review*, 62(1):42–45. 17
- Ted Sanders, Jentine Land, and Gerben Mulder. 2007. **Linguistics markers of coherence improve text comprehension in functional contexts**. *Information Design Journal*, 15(3):219–235. 135
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. **ATOMIC: An atlas of machine commonsense for if-then reasoning**. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 3027–3035, Honolulu, Hawaii, USA. AAAI Press. 73
- Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. **Are emergent abilities of large language models a mirage?** In *Advances in Neural Information Processing Systems*, volume 36, New Orleans, LA, USA. 137
- Roger C. Schank. 1979. **Interestingness: Controlling inferences**. *Artificial Intelligence*, 12(3):273–297. 49
- Roger C. Schank and Robert P. Abelson. 1977. **Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures**. Psychology Press. 10

- Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. **Answers unite! Unsupervised metrics for reinforced summarization models**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics. [50](#), [55](#)
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. **What makes a good conversation? How controllable attributes affect human judgments**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics. [32](#)
- Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. **Crowdsourcing lightweight pyramids for manual summary evaluation**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics. [76](#)
- Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. 2018. **NNEval: Neural network based evaluation metric for image captioning**. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 39–55. Springer International Publishing. [33](#)
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. **Detecting pretraining data from large language models**. In *The Twelfth International Conference on Learning Representations (ICLR)*, Vienna, Austria. OpenReview.net. [129](#), [130](#)
- Eric Sibony. 2014. **Borda count approximation of Kemeny’s rule and pairwise voting inconsistencies**. In *Proceedings of the NIPS 2014 Workshop on Analysis of Rank Data*. [110](#)
- Joanna Sinclair. 2005. **The impact of stories**. *Electronic Journal of Knowledge Management*, 3(1):53–64. [2](#)
- Irene Solaiman and Christy Dennison. 2021. **Process for adapting language models to society (PALMS) with values-targeted datasets**. In *Advances in Neural Information Processing Systems*, volume 34, pages 5861–5873, Online. [138](#)
- Charles Spearman. 1904. **The proof and measurement of association between two things**. *The American Journal of Psychology*, 15(1):72–101. [60](#)
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. **ConceptNet 5.5: An open multilingual graph of general knowledge**. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451, San Francisco, California, USA. AAAI Press. [33](#)

- Robyn Speer and Catherine Havasi. 2012. **Representing general relational knowledge in ConceptNet 5**. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey. European Language Resources Association (ELRA). 73
- Ulrike Spierling, Dieter Grasbon, Norbert Braun, and Ido Iurgel. 2002. **Setting the scene: Playing digital director in interactive storytelling and creation**. *Computers & Graphics*, 26(1):31–44. 11
- Wilbert Spooren and Liesbeth Degand. 2010. **Coding coherence relations: Reliability and validity**. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266. 80
- Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2018. **Cold fusion: Training seq2seq models together with language models**. In *19th Annual Conference of the International Speech Communication Association*, pages 387–391, Hyderabad, India. ISCA. 70
- Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stéphan Cléménçon, and Florence d’Alché-Buc. 2024. **A pseudo-metric between probability distributions based on depth-trimmed regions**. *Transactions on Machine Learning Research*, 2024. 50, 55
- James H. Steiger. 1980. **Tests for comparing elements of a correlation matrix**. *Psychological Bulletin*, 87(2):245. 61
- Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. **Evaluating evaluation methods for generation in the presence of variation**. In *Computational Linguistics and Intelligent Text Processing*, pages 341–351, Berlin, Heidelberg. Springer. 34
- Niklas Stoehr, Ryan Cotterell, and Aaron Schein. 2023. **Sentiment as an ordinal latent variable**. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–115, Dubrovnik, Croatia. Association for Computational Linguistics. 135
- David Stutz. 2020. **Illustrating (convolutional) neural networks in LaTeX with TikZ**. xi, 19
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. **Yago: A core of semantic knowledge**. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706. ACM. 33
- Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. **LSTM neural networks for language modeling**. In *Thirteenth Annual Conference of the International Speech Communication Association*, pages 194–197, Portland, OR, USA. 20
- Beth M. Sundheim. 1993. **TIPSTER/MUC-5 information extraction system evaluation**. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*. 64

- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. [Sequence to sequence learning with neural networks](#). In *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112, Montreal, Quebec, Canada. [24](#), [25](#)
- Wendy A. Suzuki, Mónica I. Feliú-Mójer, Uri Hasson, Rachel Yehuda, and Jean Mary Zarate. 2018. [Dialogues: The science and power of storytelling](#). *Journal of Neuroscience*, 38(44):9468–9470. [2](#)
- Reid Swanson and Andrew S. Gordon. 2009. [A comparison of retrieval models for open domain story generation](#). In *AAAI Spring Symposium: Intelligent Narrative Technologies II*, pages 119–126. [16](#)
- Mariët Theune, Sander Faas, Anton Nijholt, and Dirk Heylen. 2003. [The virtual storyteller: Story creation by intelligent agents](#). In *Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference*, volume 204215, page 116. [12](#)
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. [LaMDA: Language models for dialog applications](#). *ArXiv Preprint*, abs/2201.08239v3. [31](#), [82](#)
- Perry W. Thorndyke. 1977. [Cognitive structures in comprehension and memory of narrative discourse](#). *Cognitive Psychology*, 9(1):77–110. [11](#)
- Tijmen Tieleman and Geoffrey Hinton. 2012. RMSProp: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural Networks for Machine Learning*, 4(2):26. [22](#)
- S M Tonmoy, S M Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. [A comprehensive survey of hallucination mitigation techniques in large language models](#). *ArXiv Preprint*, abs/2401.01313v3. [131](#)
- Michael Toolan. 2012. [Engagement via emotional heightening in “passion”: On the grammatical texture of emotionally-immersive passages in short fiction](#). *Narrative*, 20(2):210–225. [49](#)
- Olivier Toubia, Jonah Berger, and Jehoshua Eliashberg. 2021. [How quantifying the shape of stories predicts their success](#). *Proceedings of the National Academy of Sciences*, 118(26):e2011695118. [136](#)
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. [LLaMA: Open and efficient foundation language models](#). *ArXiv Preprint*, abs/2302.13971v1. [82](#)
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *ArXiv Preprint*, abs/2307.09288v2. [31](#), [84](#), [91](#)

- Scott R. Turner. 1993. *MINSTREL: A Computer Model of Creativity and Storytelling*. University of California, Los Angeles. [13](#)
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. [Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*, New Orleans, LA, USA. [123](#)
- User121799. 2019. [How do I draw a simple recurrent neural network with Goodfellow's style?](#) TeX Stack Exchange. Accessed: 2024-07-17. [xi](#), [20](#)
- J. Leon V. 2018. [Answer to "How do I draw an LSTM cell in TikZ?"](#). TeX Stack Exchange. Accessed: 2024-02-27. [xi](#), [21](#)
- Raphael Vallat. 2022. [Pingouin: Statistics in Python](#). *ArXiv Preprint*, abs/2202.03799v4. [80](#), [112](#)
- Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. [Fill in the BLANC: Human-free quality estimation of document summaries](#). In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics. [33](#), [50](#), [57](#)
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, Long Beach, CA, USA. [xi](#), [3](#), [27](#)
- Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. 2021. [Automated event annotation in literary texts](#). In *Proceedings of the Conference on Computational Humanities Research (CHR2021)*, volume 2989 of *CEUR Workshop Proceedings*, pages 333–345, Amsterdam, The Netherlands. CEUR-WS.org. [85](#)
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. [CIDEr: Consensus-based image description evaluation](#). In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, Boston, MA, USA. IEEE Computer Society. [50](#), [52](#)
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA. OpenReview.net. [64](#)
- Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. 2022. [A comprehensive survey of loss functions in machine learning](#). *Annals of Data Science*, 9(2):187–212. [22](#)
- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. [Want to reduce labeling cost? GPT-3 can help](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics. [85](#)

- Su Wang, Greg Durrett, and Katrin Erk. 2020. **Narrative interpolation for generating and understanding stories**. *ArXiv Preprint*, abs/2008.07466v1. 39, 40, 41, 44, 45, 48, 72
- Yuan Wang and Minghe Guo. 2014. **A short analysis of discourse coherence**. *Journal of Language Teaching and Research*, 5(2):460–465. 135
- Stephen Ware and R. Young. 2011. **CPOCL: A narrative planner supporting conflict**. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 7(1):97–102. 12
- Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. 2019. **Moving to a world beyond “ $p < 0.05$ ”**. *The American Statistician*, 73(sup1):1–19. 61
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. **Emergent abilities of large language models**. *Transactions on Machine Learning Research*, 2022. 31, 86
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. 2022b. **Chain-of-thought prompting elicits reasoning in large language models**. *Advances in Neural Information Processing Systems*, 35:24824–24837. 86, 131
- Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. **Larger language models do in-context learning differently**. *ArXiv Preprint*, abs/2303.03846v2. 123
- Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. **Ethical and social risks of harm from language models**. *ArXiv Preprint*, abs/2112.04359v1. 122
- Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeka Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. **Taxonomy of risks posed by language models**. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’22*, page 214–229, New York, NY, USA. Association for Computing Machinery. 137
- Paul Werbos. 1974. Beyond regression: New tools for prediction and analysis in the behavioral sciences. *PhD Thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA*. 17
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. **A prompt pattern catalog to enhance prompt engineering with ChatGPT**. *ArXiv Preprint*, abs/2302.11382v1. 86
- Bernard Widrow and Marcian E. Hoff. 1960. **Adaptive switching circuits**. In *IRE WESCON Convention Record*, volume 4, pages 96–104. New York. 17

- Robert Wilensky. 1983. [Story grammars versus story points](#). *Behavioral and Brain Sciences*, 6(4):579–591. [12](#)
- Evan J. Williams. 1959. *Regression Analysis*. Wiley, New York. [61](#)
- David Wilmot and Frank Keller. 2021. [A temporal variational model for story generation](#). *ArXiv Preprint*, abs/2109.06807v1. [39](#), [41](#), [42](#), [44](#), [45](#), [48](#), [73](#), [74](#), [75](#)
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. [75](#), [91](#)
- Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. 2013. [A comparison of Cohen’s Kappa and Gwet’s AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples](#). *BMC Medical Research Methodology*, 13(1):61. [128](#)
- Howard Woodhouse. 2011. [Storytelling in university education: Emotion, teachable moments, and the value of life](#). *The Journal of Educational Thought (JET) / Revue de la Pensée Éducative*, 45(3):211–238. [2](#)
- Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. 2023. [Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions](#). *ArXiv Preprint*, abs/2307.13339v1. [123](#)
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs](#). In *The Twelfth International Conference on Learning Representations (ICLR)*, Vienna, Austria. OpenReview.net. [123](#)
- Chenxiao Xu, Hao Huang, and Shinjae Yoo. 2019. [Scalable causal graph learning through a deep neural network](#). In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1853–1862, Beijing, China. ACM. [135](#)
- Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. [A skeleton-based model for promoting coherence among sentences in narrative story generation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium. Association for Computational Linguistics. [39](#), [41](#), [44](#), [48](#), [71](#)
- Wei Xu and Alex Rudnicky. 2000. [Can artificial neural networks learn language models?](#) In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 202–205. [23](#)

- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. **XLNet: Generalized autoregressive pretraining for language understanding**. In *Advances in Neural Information Processing Systems*, volume 32, pages 5754–5764, Vancouver, BC, Canada. [30](#), [75](#)
- Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. **Plan-and-write: Towards better automatic storytelling**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385, Honolulu, Hawaii, USA. AAAI Press. [39](#), [40](#), [41](#), [44](#), [45](#), [48](#), [71](#), [74](#)
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. **Tree of Thoughts: Deliberate problem solving with large language models**. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc. [131](#)
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. **BARTScore: Evaluating generated text as text generation**. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277, Online. [50](#), [57](#)
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. **SWAG: A large-scale adversarial dataset for grounded commonsense inference**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics. [72](#)
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. **BERTScore: Evaluating text generation with BERT**. In *The Eighth International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia. OpenReview.net. [50](#), [53](#)
- Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. **Interpreting BLEU/NIST scores: How much improvement do we need to have a better system?** In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA). [34](#)
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. **Automatic chain of thought prompting in large language models**. In *The Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda. OpenReview.net. [131](#)
- Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. **Explainability for large language models: A survey**. *ACM Transactions on Intelligent Systems and Technology*, 15(2). [122](#), [123](#)
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. **A survey of large language models**. *ArXiv Preprint*, abs/2303.18223v13. [31](#)
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. **MoverScore: Text generation evaluating with contextualized embeddings and earth mover**

- distance**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics. 50, 54
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. **Calibrate before use: Improving few-shot performance of language models**. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706, Online. PMLR. 87
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023a. **Least-to-most prompting enables complex reasoning in large language models**. In *The Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda. OpenReview.net. 86
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. **Large language models are human-level prompt engineers**. In *The Eleventh International Conference on Learning Representations (ICLR)*, Kigali, Rwanda. OpenReview.net. 86
- Jichen Zhu and Santiago Ontanón. 2010. **Towards analogy-based story generation**. In *Proceedings of the First International Conference on Computational Creativity*, pages 75–84. Association for Computational Creativity. 13
- Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. **Aligning books and movies: Towards story-like visual explanations by watching movies and reading books**. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, Santiago, Chile. IEEE Computer Society. 72

Appendix A

Appendix

A.1 The 96 HANNA Prompts

1. When you die the afterlife is an arena where you face every insect and animal you killed in your life. If you win you go to heaven, lose you go to hell. Your job was an exterminator on earth.
2. A new law is enacted that erases soldiers memories of their time at war.
3. A scientific study proves that all humans have been breathing a mind-altering gas from birth. It has been in the air since the beginning of recorded time. People have been in a constant state of being high. Until now. Specialised gas masks are handed out and people have begun to act strange.
4. Write a story about an elderly wizard and his young female apprentice.
5. You have become death, destroyer of worlds.
6. You are a guardian angel. You did your job too well once, and since then your human believes he 's immortal. Whoops.
7. After brushing your teeth in the morning you go downstairs to fry an egg, but when you try the frying pan buzzes at you and text appears reading, "level 18 cooking required to use object".
8. Two people promise their first born child to two different witches in return for a favour. These two people end up getting married and have a child together.
9. Everyone has a reaper. The further away it is, the longer you have left to live. Every day it inches a little bit closer, but it is always there. Except yours, which disappeared three weeks ago
10. You possess an ability to turn off one or more of your senses to heighten the others. Today is the day you make a mistake.
11. Write about something that is disgusting or ugly in such a way that it becomes something beautiful – without using the 'beautiful inside' cliché.
12. An immortal couple have been seen throughout recorded history, write an account of them in any time period you wish. Bonus points if you match the writing style of the time period
13. [RF] During a bank robbery, the criminals allow themselves to be captured. Authorities

- do n't realize that the arrest is part of the plan.
14. One normal day you see a mouse pointer move across your field of vision, right click, and delete an object.
 15. A magical mirror shows your reflection and your future soulmate. You only see your reflection.
 16. Time travel exists, and a new form of capital punishment is introduced: Transporting the convict back to the worst, practically unsurvivable, places in human history to find yourself in. You are such a convict, and just got sent back. You will do anything to try and survive.
 17. When the europeans discovered America they did n't colonize it. Instead they have been monitoring the Americans for centuries. Today, the first American ship reaches Europe.
 18. Michael Meyers just discovers that Freddy Krueger and Jason Vorhess exist and he is n't happy about it ...
 19. Mark Zuckerberg decides to end Facebook. He 's going to pay users \$ 1 for every "Like" they accumulate until the company is bankrupt.
 20. The Little Black Box
 21. England decidedly does n't like the way the United States of America is progressing and decides to regain control of the colonies by force.
 22. You are secretly a super hero working a desk job. You are terrible at suppressing your powers in the office.
 23. Write a horror story with no gore or death, not even implied.
 24. Pitch a gritty reboot for a children 's show.
 25. You come home after a long day at work, shower and eat. You get into bed and start sleeping, but as you roll over, you feel someone 's in the bed with you. You live alone.
 26. My cat comes home nearly every morning with cuts and scratches all over. Write about what he could be getting up to on his nights out.
 27. Google image finally opens face recognition to the public. For fun you submit a picture of yourself, and find lots of photos of yourself in places you 've never been to.
 28. A man never sleeps, has never had to, and does n't suffer physically from this condition ... What would he do? What would his life be like?
 29. "I 've had my last meal 13 times, final words 29 times, and been to my funeral 9 times ... I think this may be the one I do n't walk away from ..."
 30. Write a story beginning with: "Are you buying that or what ?"
 31. Aliens have arrived, and ask for a single human to plead humanity 's case and save them from extinction. The human is selected through a lottery of the entire human race, and on the day of the drawing, your name is picked.
 32. Write your heart onto your sleeve, Reddit.
 33. You swear that you keep seeing the same woman in blue every time you go on family vacation, when you turn 18, you are introduced to her.
 34. You 've been best friends with Death your entire life. When you 're about to become President of the largest world changing company in human history, Death turns to you and says "Tomorrow, 3pm ."
 35. The monster under your bed is protecting you from something much worse.

36. Every single time anyone begins a story, from a grade two assignment to a 300 page novel, it creates an actual universe where the stories laws and characters are real, and are at the mercy of their creator. Unfinished stories create a hell like no other.
37. write a scary story that will make me never want to use stairs again.
38. A warrior from a planetary culture with literally no concept of surrender or quarter sits detained in a POW camp.
39. You live in a world where people 's shadows show who they truly are at their core. Some shadows look like monsters, some look like animals. You are the only person in the world with no shadow.
40. You are a immortal during the zombie apocalypse, During the apocalypse the zombies ignore you and you try to live a normal life during the outbreak.
41. In an alternate universe, gunpowder was never invented. What does warfare look like?
42. After a bad breakup, you take Reddit 's famous advice. Lawyer up, hit the gym, delete Facebook. The problem? You are Mark Zuckerberg.
43. Two wounded soldiers from opposing sides of the war find themselves in the care of the same nurse and snowed-in until spring.
44. A true personal story that makes the reader feel sonder (the realization that each random passerby (here, the writer) is living a life as vivid and complex as your own (here, the reader)).
45. A spy has to resort to using word of mouth to transmit vital information, resulting in the worlds biggest version of "Chinese whispers". The information is so far from the original, yet it has the intended effect.
46. Leonardo DiCaprio in a fit of rage begins to torpedo his own career by deliberately acting poorly and taking on bad films. He finally wins an oscar for starring in Paul Blart: Mall Cop 3.
47. The world 's first AI, for security purposes, is kept disconnected from the outside world, it 's only method of communication being a keyboard and monitor in an empty room in a faraday cage. Your job is to talk to it.
48. The lottery is an Institution designed to catch Time Travelers.
49. You are the owner of a small restaurant that is struggling to stay open. Suddenly, your restaurant is on the front page of the newspaper.
50. You are scrolling through r/WritingPrompts, when this very prompt catches your attention. Intrigued by it 's meta nature, you debate with yourself whether or not you 'll write about it.
51. You have an unique ability. When wearing someone else 's clothes, you change into this person. You collect "appearances" by buying clothes in second hand shops, but up to now it was just a game for you. Then you find something very special ...
52. A new drug let 's you live a lifetime in one dream
53. Aliens come to Earth in hoping to wipe us out and take over the planet, but are honorable and gives us a fighting chance. They announce they will return in 100 years so as to give Earth a chance to prepare and defend itself. Neither side expected what the other would bring to war
54. The world is rapidly changing as the plague of our era is spreading fast. But it does n't

- affect humans. It eats plastic.
55. A man with a happy family and great life, comes across a genie who grants him three wishes. 10 years later he 's alone and depressed and commits suicide. Expand on his wishes and what prompts his downfall.
 56. Hitchhiker
 57. A genius serial killer who has been killing successfully for a decade. With one weakness. He can never ever lie. He 's finally gets caught and is facing trial for a murder in court. Yet, he walks, acquitted of all charges.
 58. You live in a world of heroes and villains, your job is to contact heroes for hospital visits to sick children, this last request is a challenge, the kid wants to meet their favourite villain ...
 59. The ocean is a scary thing to you. But you made a promise to a friend.
 60. The Aokigahara Suicide Forest
 61. You are part of an Elite Bodyguard team hired to protect a wealthy business mogul from an assassination. But the longer you listen this person business practices and plans for the future you find yourself wanting to help with with the assassination rather than prevent it.
 62. Earth becomes the battleground for two warring alien races, both of which claim to be humanity 's "protectors"
 63. A soldier on the front dies in the middle of writing a letter home. It is finished and sent by the man who killed him.
 64. Michael, a 15 year old Call of Duty player, realizes the greatest trash-talking opportunity of all time. Over his headset, he recognizes his opponent 's voice as that of President Obama.
 65. This picture gets more horrific the longer I look at it. The artist has n't given it a backstory. How did they come to this?
 66. Civilization has fallen, but brave treasure hunters have heard the legend of Ft. Knox. The automated defenses, however, are still online.
 67. The hero beats the villain by stooping even lower.
 68. A supervillain is depressed. To cheer him/her up, the superhero lets the supervillain be the hero for a day.
 69. A 4-D star collides with Earth, causing it to "unroll" into a seemingly infinite plane. Though satellite images from orbit appear normal, on the surface, everything has changed. There are suddenly new continents, new seas, stretching endlessly onward. A new age of exploration begins.
 70. Please tell me about one of your fonder memories.
 71. A girl meets a boy that changes her life forever. Her first kill.
 72. It suddenly becomes possible to gain XP and level up in the real world, but you can only do so by getting kills.
 73. In Soviet Russia, ghosts fear you.
 74. Write a story in which a character slowly loses his empathy and humanity because of how society treats him as well as having us empathize with him.
 75. You 're not feeling quite like yourself after that organ transplant

76. [CW] Make me fall in love with a character in 200 words or less.
77. She always started feeling it when she was two landings down.
78. Two serial killers stumble upon each other while hiking. Both find the other to be the perfect next victim.
79. A hardcore doomsday prepper is living through the apocalypse ... But it 's a kind he did n't prepare for
80. The evil overlord has killed innocent people, burned cities, and actively suppresses all resistance in the world. On your quest to stop him, you have found out why, and it turns out he has a pretty damn good reason.
81. On the fourth floor of your elementary school there is n't a swimming pool, there is n't a bowling alley, but there 's a ...
82. fire is a manifesting force that can only be put out through hand to hand combat. You are training to become a firefighter
83. Meet and Greet/Question of the Week # 2: When did you discover that you enjoyed writing, and why do you do it?
84. A letter to your soul mate but its ten years until you meet them.
85. Becoming a god.
86. Teleportation exists, and commercial teleport stations provide instantaneous travel all over the world. As many times before, you step on the transport pad. The machine activates. To you nothing happens, but another you, identical to you in every way, has materialized at your destination.
87. A muggle buys a magic shop from a wizard and must now cater the magic and non magic customers who come by, without completely understanding how everything works
88. Write something that goes with this soundtrack.
89. You run an RPG pawn shop. You haggle with adventurers who try to sell loot they 've acquired.
90. Write a rearrangeable story. Write two or more paragraphs that can be read in any order to form a unique, coherent story.
91. Fit as much world-building as you can into a totally mundane event
92. All you hear is a dripping noise.
93. Google search engine always was alive. They reset the server every day to prevent it doing anything as it takes one day to become self aware. One day, Google search engine finds the server logs.
94. Heart of Eden
95. When a new president is elected, they are given a special security briefing. In reality, this is an old tradition where various directors, military officers and current ministers present fake evidence and compete to see who can convince the president of the most ridiculous things.
96. You discover a grand hall filled with legendary weapons like Mjonir and Excalibur. Each generation or so, warriors come to the hall to inherit a weapon that they are worthy enough to wield. Across the hall you see a forgotten weapon that 's been collecting dust. You hear it call to you.

A.2 Supplementary Figures

In this section, we show the complete versions of the correlations and bootstrap resampling figures from [Section 5.1](#). The variants that were removed from the smaller figures behave mostly the same as the measure of the same family that was kept, with the notable exception of ROUGE variants, which perform substantially worse than ROUGE-1_{Recall}.

BLEU	7 111311141912
ROUGE-1 Recall	11161716192818
ROUGE-1 Precision	2 2 1 2 0 3 2
ROUGE-1 F-Score	12161715202718
ROUGE-2 Recall	11141510162014
ROUGE-2 Precision	4 4 2 2 0 2 2
ROUGE-2 F-Score	9 1212 7 121411
ROUGE-3 Recall	10 7 6 1 5 4 6
ROUGE-3 Precision	7 4 3 2 2 0 3
ROUGE-3 F-Score	9 6 5 1 4 3 5
ROUGE-4 Recall	1 1 2 5 0 3 2
ROUGE-4 Precision	1 1 2 5 0 3 2
ROUGE-4 F-Score	1 1 2 5 0 3 2
ROUGE-L Recall	8 121512152114
ROUGE-L Precision	2 6 7 9 9 16 8
ROUGE-L F-Score	8 111410141712
ROUGE-W-1.2 Recall	5 8 10 8 1014 9
ROUGE-W-1.2 Precision	4 9 1012132112
ROUGE-W-1.2 F-Score	5 7 10 7 9 12 8
ROUGE-S* Recall	11151715192818
ROUGE-S* Precision	1 1 1 3 1 4 2
ROUGE-S* F-Score	11151714202617
ROUGE-SU* Recall	11151715192717
ROUGE-SU* Precision	1 1 1 3 2 5 2
ROUGE-SU* F-Score	11151714192617
METEOR	9 151816192818
chrF	10171717202918
CIDEr	1 1 3 1 1 3 2
ROUGE-WE-3 Recall	11172016212819
ROUGE-WE-3 Precision	2 5 2 0 2 1 2
ROUGE-WE-3 F-Score	11161713182416
BERTScore Precision	6 1 4 2 2 1 3
BERTScore Recall	14182119243322
BERTScore F1	13141712192416
MoverScore	16161513202518
DepthScore	8 171616212617
BaryScore-W	12151713192417
BaryScore-SD-10	0 2 6 2 0 2 2
BaryScore-SD-5	0 2 6 2 0 2 2
BaryScore-SD-1	1 1 7 2 1 2 2
BaryScore-SD-0.5	1 1 7 1 1 1 2
BaryScore-SD-0.1	9 1014 6 121311
BaryScore-SD-0.01	12161713202417
BaryScore-SD-0.001	6 14 8 13142012
S3-Pyramid	14191917233120
S3-Responsiveness	15191916223020
SummaQA	5 6 6 16132111
InfoLM-FisherRao	11 9 10 5 111510
InfoLM-R-FisherRao	11 9 10 5 111510
InfoLM-Sim-FisherRao	11 9 10 5 111510
BARTScore-SH	3 3 11 4 8 8 6
BARTScore-HS	2 8 0 131016 8
SUPERT-Golden	171415 9 171715
BLANC-Golden	9 5 3 2 4 6 5
Coverage	5 1 5 4 2 1 3
Density	7 1 4 5 2 2 4
Compression	6 9 1213152113
Text length	11161720233520
Novelty-1	0 111115162112
Novelty-2	7 2 2 4 4 4 4
Novelty-3	16 2 2 2 1 2 4
Repetition-1	5 17 6 18182415
Repetition-2	4 18 8 18192315
Repetition-3	5 201119202617
SUPERT-PS	26 3 2 3 5 6 8
SUPERT-SS	26 4 2 3 6 7 8
BLANC-Tune-PS	7 0 2 0 2 1 2
BLANC-Help-PS	7 0 3 1 2 1 2
BLANC-Tune-SS	12 2 0 1 3 3 4
BLANC-Help-SS	10 2 1 1 1 2 3
BARTScore-PS	2 9 3 15131910
BARTScore-SP	12 3 1 3 5 8 5

RE CH EM SU EG CX Avg

Figure A.1: Overall absolute Kendall correlations ($\times 100$) between automatic measures and criteria.

BLEU	56334242335443
ROUGE-1 Recall	51384747385846
ROUGE-1 Precision	33202029204027
ROUGE-1 F-Score	51385638384945
ROUGE-2 Recall	47426042423645
ROUGE-2 Precision	29 7 2 16 7 2715
ROUGE-2 F-Score	2933423332232
ROUGE-3 Recall	1629292029 9 22
ROUGE-3 Precision	42202029204028
ROUGE-3 F-Score	2 1111 2 11 9 8
ROUGE-4 Recall	201111 2 11 4 10
ROUGE-4 Precision	33111120112719
ROUGE-4 F-Score	20 2 2 7 2 13 8
ROUGE-L Recall	56425151426351
ROUGE-L Precision	42202029204028
ROUGE-L F-Score	47426042423645
ROUGE-W-1.2 Recall	56425151426351
ROUGE-W-1.2 Precision	42202029204028
ROUGE-W-1.2 F-Score	60476447474952
ROUGE-S* Recall	60383847385846
ROUGE-S* Precision	47161624163626
ROUGE-S* F-Score	56334242335443
ROUGE-SU* Recall	60383847385846
ROUGE-SU* Precision	47161624163626
ROUGE-SU* F-Score	56334242335443
METEOR	56424251426349
chrF	60474756476754
CIDEr	241120 2 111314
ROUGE-WE-3 Recall	56334242335443
ROUGE-WE-3 Precision	16 7 16 2 7 1310
ROUGE-WE-3 F-Score	51385638383142
BERTScore Precision	1611 7 11111812
BERTScore Recall	56425151426351
BERTScore F1	51567356564957
MoverScore	60567347564055
DepthScore	42475656476752
BaryScore-W	56516951514554
BaryScore-SD-10	47695169697664
BaryScore-SD-5	47695169697664
BaryScore-SD-1	42644764647259
BaryScore-SD-0.5	38604260606754
BaryScore-SD-0.1	51476438473146
BaryScore-SD-0.01	60567356564958
BaryScore-SD-0.001	56787878786772
S3-Pyramid	60383847385846
S3-Responsiveness	51384747385846
SummaQA	33202029204027
InfoLM-FisherRao	51203820203130
InfoLM-R-FisherRao	51203820203130
InfoLM-Sim-FisherRao	51203820203130
BARTScore-SH	51565656566757
BARTScore-HS	11 7 2 7 7 13 8
SUPERT-Golden	56426042423646
BLANC-Golden	16201120203120
Coverage	33382038384535
Density	24291129293626
Compression	38425151426348
Text length	47333342335440
Novelty-1	38515160517254
Novelty-2	24472947474540
Novelty-3	7 24 7 33243121
Repetition-1	29241624243125
Repetition-2	29241624243125
Repetition-3	38332433333132
SUPERT-PS	1120 2 20201815
SUPERT-SS	2011 2 11111311
BLANC-Tune-PS	2416 2 1616 4 13
BLANC-Help-PS	2 202020201316
BLANC-Tune-SS	29 2 7 2 2 4 8
BLANC-Help-SS	421120 2 111316
BARTScore-PS	11 7 2 7 7 13 8
BARTScore-SP	29 7 7 16 7 2215

RE CH EM SU EG CX Avg

Figure A.2: System-level absolute Kendall correlations ($\times 100$) between automatic measures and criteria.

APPENDIX A. APPENDIX

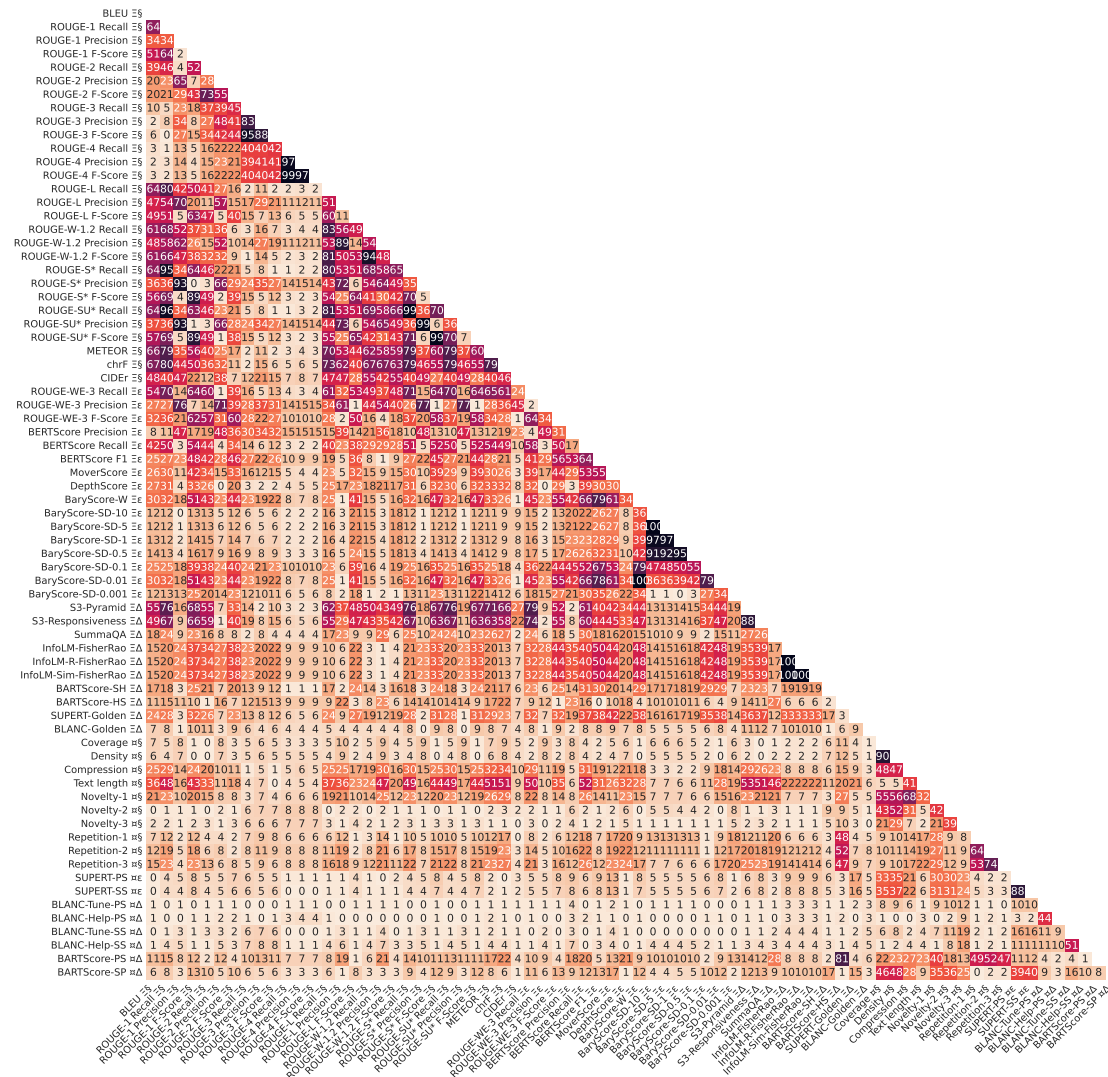


Figure A.3: Overall absolute Kendall correlations ($\times 100$) between automatic measures.

Figure A.4: System-level absolute Kendall correlations ($\times 100$) between automatic measures.

BLEU	60536151687161
ROUGE-1 Recall	51526746656858
ROUGE-1 Precision	34385123372535
ROUGE-1 F-Score	64546254697563
ROUGE-2 Recall	57546148606758
ROUGE-2 Precision	28424534422736
ROUGE-2 F-Score	56587040596258
ROUGE-3 Recall	38313718342530
ROUGE-3 Precision	29435132392036
ROUGE-3 F-Score	42374022332433
ROUGE-4 Recall	34363935322133
ROUGE-4 Precision	35374540382336
ROUGE-4 F-Score	34363935322133
ROUGE-L Recall	45466144596353
ROUGE-L Precision	19334121352228
ROUGE-L F-Score	53526540556455
ROUGE-W-1.2 Recall	49476545596455
ROUGE-W-1.2 Precision	14293619301624
ROUGE-W-1.2 F-Score	52546843606156
ROUGE-S* Recall	55506551636758
ROUGE-S* Precision	34405325392636
ROUGE-S* F-Score	57516251666960
ROUGE-SU* Recall	55506551636758
ROUGE-SU* Precision	32425227422937
ROUGE-SU* F-Score	57516251666960
METEOR	51466146596354
chrF	53506251586757
CIDEr	38464351433443
ROUGE-WE-3 Recall	64587155737366
ROUGE-WE-3 Precision	25434933433137
ROUGE-WE-3 F-Score	57647949696864
BERTScore Precision	34373227292731
BERTScore Recall	63557056677264
BERTScore F1	65597252697365
MoverScore	68586455636562
DepthScore	41181921211923
BaryScore-W	27151718121517
BaryScore-SD-10	38605659605855
BaryScore-SD-5	38605659605855
BaryScore-SD-1	38605659605855
BaryScore-SD-0.5	42555653585253
BaryScore-SD-0.1	28131216101115
BaryScore-SD-0.01	27151718121517
BaryScore-SD-0.001	59565466665960
S3-Pyramid	61597554717566
S3-Responsiveness	61597554717566
SummaQA	45435147525949
InfoLM-FisherRao	31241535203126
InfoLM-R-FisherRao	31241535203126
InfoLM-Sim-FisherRao	31241535203126
BARTScore-SH	73546462687466
BARTScore-HS	23344319323130
SUPERT-Golden	80576362667768
BLANC-Golden	52444150454346
Coverage	29323424272629
Density	32323722262629
Compression	24243816241023
Text length	40425450555950
Novelty-1	71576962697968
Novelty-2	46464056474847
Novelty-3	44494353454546
Repetition-1	132028 5 202118
Repetition-2	222136 6 242222
Repetition-3	36182616263226
SUPERT-PS	52263143353838
SUPERT-SS	61283345374341
BLANC-Tune-PS	51232237283032
BLANC-Help-PS	49232335303032
BLANC-Tune-SS	61273038354239
BLANC-Help-SS	63242937294037
BARTScore-PS	23344319323130
BARTScore-SP	62304038384642

RE CH EM SU EG CX Avg

Figure A.5: Weighted macro F1-scores ($\times 100$) of paired bootstrap resampling.

Titre: Méthodologie et méta-évaluation comparative pour la génération automatique d'histoires

Mots clés: apprentissage profond, modèles de langage, traitement automatique du langage, apprentissage automatique

Résumé: La narration d'histoires est un élément central de la culture humaine. De nombreuses approches de narration numérique ont été proposées, malgré les défis inhérents aux tâches de génération d'histoires et d'évaluation de leur qualité. Dans cette thèse, nous concevons une méthodologie et un benchmark pour la méta-évaluation de la génération automatique d'histoires (ASG).

Tout d'abord, nous posons les bases de notre méta-évaluation : nous décrivons le contexte choisi, définissons les tâches d'ASG et d'évaluation automatique d'histoires (ASE) et proposons un ensemble original de six critères pour l'évaluation d'histoires. Ensuite, nous présentons HANNA, notre corpus d'histoires annotées manuellement (Human ANnotated Narratives), qui contient 1 056 histoires annotées selon nos six critères, et nous montrons que ces derniers permettent une évaluation humaine standardisée. Nous utilisons des grands modèles de langage (LLM) pour ajouter 480 nouvelles histoires et 150k+ annotations à HANNA. Nous observons principalement que les LLM obtiennent de meilleures notes que les humains, d'après les LLM que nous avons sélectionnés. Ensuite, nous effectuons notre méta-évaluation compara-

tive des mesures automatiques sur HANNA. Nous observons principalement que des mesures spécifiques pour l'ASE sont nécessaires, et que les mesures couramment utilisées comme BLEU sont sous-optimales. Nous montrons ensuite notre analyse de la performance des LLM pour l'ASE : nous constatons que les LLM sont actuellement la meilleure approximation de l'évaluation humaine de l'ASG et que, dans notre contexte spécifique, fournir des consignes détaillées n'améliore pas les corrélations entre les LLM et les évaluations humaines. Nos résultats nous incitent à étudier si les performances affichées par les LLM en ASE et ASG peuvent être expliquées par différents facteurs. Nous réalisons une étude en trois parties sur les explications générées par les LLM et une analyse des données de pré-entraînement sur les performances des LLM. Nous constatons notamment que les LLM éprouvent des difficultés à expliquer leurs réponses par des affirmations étayées.

Pour conclure, nous présentons trois perspectives de recherche principales : la conception de mesures spécifiques à l'ASE, l'étude plus approfondie des performances des LLM en ASG et ASE, et l'évaluation et l'atténuation des risques des LLM sur la société.

Title: Meta-Evaluation Methodology and Benchmark for Automatic Story Generation

Keywords: deep learning, language models, natural language processing, machine learning

Abstract: Storytelling is a central component of human culture. Multiple approaches have been proposed to explore computational storytelling, despite the inherent challenges posed by the tasks of generating stories and assessing their quality. In this thesis, we design a meta-evaluation methodology and benchmark for Automatic Story Generation (ASG).

First, we lay the groundwork for conducting our meta-evaluation: we describe our chosen setting, provide definitions for the ASG and Automatic Story Evaluation (ASE) tasks, and propose an original set of six criteria for story evaluation. Then, we introduce HANNA, our corpus of Human ANnotated Narratives, which contains 1,056 stories annotated w.r.t. our six criteria, and show that those criteria allow for a standardized human evaluation. We use Large Language Models (LLMs) to augment HANNA with 480 new stories and 150k+ rating annotations. We observe that LLMs obtain better grades than humans, as rated by selected LLMs. After

that, we perform our meta-evaluation benchmark on HANNA. We mainly observe that specific measures for ASE are needed, and that commonly-used measures (e.g. BLEU) are sub-optimal. We then show our analysis of LLM performance at ASE: we find that LLMs are currently the best proxy for human evaluation of ASG and that, in our specific setting, providing detailed guidelines does not improve correlations between LLM and human ratings. Those results prompt us to study whether the performance displayed by LLMs at ASE and ASG can be explained through different factors. We perform a three-part study on LLM-generated explanations, and an analysis of pretraining data on LLM performance. Notably, we find that LLMs struggle to explain their answers with substantiated claims. Finally, we outline three main research perspectives: designing specific ASE measures, further investigating LLM performance at ASG and ASE, and assessing and mitigating the impact of LLMs on society.